

Adaptive 3D Facial Action Intensity Estimation and Emotion Recognition

Yang Zhang¹, Li Zhang¹ and Mohammed Alamgir Hossain²

¹Computational Intelligence Research Group
Department of Computer Science and Digital Technologies
Faculty of Engineering and Environment
Northumbria University
Newcastle, NE1 8ST, UK

²Anglia Ruskin IT Research Institute, Faculty of Science and Technology,
Anglia Ruskin University
Cambridge, CB1 1PF, UK

Abstract: Automatic recognition of facial emotion has been widely studied for various computer vision tasks (e.g. health monitoring, driver state surveillance and personalized learning). Most existing facial emotion recognition systems, however, either have not fully considered subject-independent dynamic features or were limited to 2D models, thus are not robust enough for real-life recognition tasks with subject variation, head movement and illumination change. Moreover, there is also lack of systematic research on effective newly arrived novel emotion class detection. To address these challenges, we present a real-time 3D facial Action Unit (AU) intensity estimation and emotion recognition system. It automatically selects 16 motion-based facial feature sets using minimal-redundancy-maximal-relevance criterion based optimization and estimates the intensities of 16 diagnostic AUs using feedforward Neural Networks and Support Vector Regressors. We also propose a set of six novel adaptive ensemble classifiers for robust classification of the six basic emotions and the detection of newly arrived unseen novel emotion classes (emotions that are not included in the training set). A distance-based clustering and uncertainty measures of the base classifiers within each ensemble model are used to inform the novel class detection. Evaluated with the Bosphorus 3D database, the system has achieved the best performance of 0.071 overall Mean Squared Error (MSE) for AU intensity estimation using Support Vector Regressors, and 92.2% average accuracy for the recognition of the six basic emotions using the proposed ensemble classifiers. In comparison with other related work, our research outperforms other state-of-the-art research on 3D facial emotion recognition for the Bosphorus database. Moreover, in on-line real-time evaluation with real human subjects, the proposed system also shows superior real-time performance with 84% recognition accuracy and great flexibility and adaptation for newly arrived novel (e.g. ‘contempt’ which is not included in the six basic emotions) emotion detection.

Keywords: Facial Emotion Recognition, Action Unit Intensity Estimation, Adaptive Ensemble Classifiers, Complementary Neural Networks, Support Vector Regression, and Support Vector Classification.

1 INTRODUCTION

Facial expressions play important roles in indicating people’s intentions, feelings and other internal states. The existing research on automatic perception of human emotions not only opened up a new era for Human-Computer Interaction research, but also showed great potential to benefit a wide variety of applications, such as computer assisted learning (D’Mello & Graesser, 2010), driver state surveillance (Vural et al., 2008), health monitoring (Lucey et al., 2009), anomalous event detection (Ryan et al., 2009), and interactive computer games (G’Mussel & Hewig, 2013).

Moreover, Facial Action Coding System (FACS) (Ekman et al., 2002) is widely used for facial emotion research in both psychology and computer science fields. It is an objective and comprehensive system based on the research of experimental

psychologists, which aims to provide human expert observers with objective measures of facial activities. In the field of behavioral science, FACS represents the most recognized standard for facial emotion measurement. A total of 46 facial Action Units (AUs) is defined to represent all possible subtle changes in muscle activations caused by emotional expressions, conversational and other facial behaviors. The original coding rules are generated based on visually discernible facial appearance changes observed from a large amount of images. According to FACS, every facial expression can be decomposed and represented by one AU or a combination of AUs. The intensity of an AU can be scored on a five-point ordinal level, from A to E (see

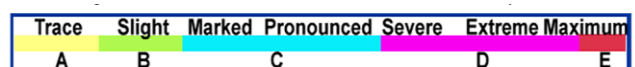


Figure 1. The five levels for AU intensity scores (Ekman et al., 2002)








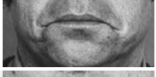




Figure 1). The definitions of these levels are provided in the following. Level A refers to a trace of an action. Level B indicates slight evidence. Level C describes pronounced or marked evidence. Level D represents severe or extreme actions with Level E indicating maximum evidence. Each intensity level refers to a range of appearance changes.

However, due to the subtleness of facial expressions and extensive coding rules defined in FACS, the AU annotation is a tedious and time consuming task and requires certified human annotators. Thus, automatic AU intensity estimation as well as emotion recognition have drawn increasing attention. The last decade has witnessed significant progress in the related areas (e.g. Cohn et al., 2009; Bartlett et al., 2006; Tian, 2002; Wen & Huang, 2003; Sorci & Thiran, 2010; Kappas, 2010; Pantic & Patras, 2006; Tsalakanidou & Malassiotis, 2010; Zhang, 2011; Valstar & Pantic, 2012; Chang et al., 2004; Koelstra et al., 2010; Antonini et al., 2006; Savran et al., 2012; Wang et al., 2006; Mpiperis, 2008; Zhang et al., 2013; Owusu et al., 2014; Rao et al., 2011). Currently, a number of systems have been developed to detect six basic emotions and their most associative AUs from images or video sequences. Many existing systems, however, either only considered static facial features, or were limited to 2D facial models. Therefore, such systems tend to lose dynamic information of facial movements that may play a critical role in interpreting emotion, and are often not robust enough against subject variation and illumination changes. Moreover, a good facial emotion recognition system is also expected to be well capable of detecting the arrival of novel emotion classes (e.g. compound emotions or other new emotions that do not belong to the six basic emotion categories mentioned in the training set). However, there is lack of systematic research for the effective detection of novel emotions.

In this paper, we present a fully automatic system for real-life 3D AU intensity estimation and facial emotion recognition. An automatic feature selection optimization algorithm is proposed to extract dynamic motion-based facial features. Neural Networks and Support Vector Regressors are then used to estimate the intensities of 16 selected Action Units with the corresponding automatically selected feature set for each AU as inputs. We also propose a set of six novel adaptive ensemble classifiers for robust recognition of the six basic emotions (i.e. happiness, surprise, fear, anger, sadness, and disgust (Ekman & Friesen, 1971)) and novel emotion detection. This research has the following distinctive contributions:

1. We extract dynamic motion-based facial features (e.g. the elongation of mouth) rather than static features (e.g. the width of mouth) to estimate AU intensities because of the following. Static features could change a lot between different subjects, whereas the motion-based features are caused by underlying facial muscle movements which bear anatomically similar muscle tension behavior among different subjects for the expression of the six basic emotions [Ekman et al., 2002], and thus are relatively universal and subject-independent, and contain comparatively richer emotional information. Therefore they are employed in this research for facial representations.
2. An automatic feature selection method based on minimal-redundancy-maximal-relevance criterion (mRMR) is proposed to identify the most discriminative and informative feature sets for AU intensity estimation. Compared with

TABLE 1
AUS, ASSOCIATED FACIAL MUSCLES, AND CORRESPONDING EXPRESSIONS (EKMAN ET AL., 2002)

<i>AU Number and Name</i>	<i>Facial Muscles</i>	<i>possible expressions</i>	<i>samples</i>
<i>AU1 Inner Brow Raiser</i>	Frontalis, Pars Medialis	Sadness	
<i>AU2 Outer Brow Raiser</i>	Frontalis, Pars Lateralis	Anger, Surprise	
<i>AU4 Brow Lowerer</i>	Procerus	Anger, Anxiety, Pain	
<i>AU5 Upper Lid Raiser</i>	Levator Palpebrae Superioris	Fear, Surprise, Anger	
<i>AU6 Cheek Raiser</i>	Orbicularis Oculi, Pars Orbitalis	Happiness	
<i>AU10 Upper Lip Raiser</i>	Levator Labii Superioris	Disgust	
<i>AU12 Lip Corner Puller</i>	Zygomaticus Major	Happiness	
<i>AU15 Lip Corner Depressor</i>	Triangularis	Sadness, Unsatisfying	
<i>AU20 Lip Stretcher</i>	Risorius	Fear	
<i>AU23 Lip Tightener</i>	Orbicularis Oris	Anger (Very)	
<i>AU24 Lip Pressor</i>	Orbicularis Oris	Anxiety	
<i>AU26 Jaw Drop</i>	Maseter	Surprise	

the manual feature selection conducted based on facial muscle anatomical and FACS knowledge, the mRMR-based optimization yields comparable performance for the intensity estimation of the 16 selected AUs.

3. We also propose a set of six novel adaptive ensemble classifiers to robustly differentiate between the six basic emotions and identify newly arrived unseen novel emotion categories. Each ensemble model employs a special type of Neural Network, i.e. Complementary Neural Network, as the base classifier, which is able to provide uncertainty measure of its classification performance. We consider the following idea for novel class detection. Instances within the same emotion categories should be close to each other whereas those from different categories should indicate great distinction to each other. Therefore, a distance-based clustering and the uncertainty measures of the base Complementary Neural Network classifiers are used to inform the arrival of novel unseen emotion classes. The proposed ensemble models achieve 92.2% average accuracy and consistently outperform other single Support Vector Machine classifiers employed in this research and other related research reported in the literature when evaluated with the Bosphorus database.

4. The proposed system is also evaluated with real-time emotion detection tasks contributed by real human subjects. The system achieves comparable accuracy (84%) in comparison to the results gained from the evaluation using database images. It also shows great adaptation and robustness for newly arrived novel emotion class detection with $\geq 70\%$ accuracy. The system is therefore proved to be effective in dealing with challenging real-life emotion recognition tasks.

The rest of the paper is organized as follows. Section 2 introduces Facial Action Coding System and discusses existing work in the related fields. We describe the methodology and implementation of the system, including facial geometric feature tracking, mRMR-based feature selection, AU intensity estimation and facial emotion recognition, in Section 3. The experiments and both on-line and off-line evaluations for AU intensity estimation and emotion recognition are discussed in Section 4. Finally, we draw conclusions and identify future work in Section 5.

2 RELATED WORK

In this section, we first of all introduce some essential FACS domain knowledge. We then discuss existing research work in the related field and conduct a concise survey on representative developments.

2.1 FACS and Related Facial Muscle Anatomy

In the Facial Action Coding System, a total of 46 unique Action Units, which are anatomically related to the contraction and relaxation of one or a specific set of facial muscles, is defined. There are 17 facial muscles, which attach to each other or to facial skin. They are innervated by facial nerve, and generate every subtle change of Action Units and facial expressions.

Moreover, according to FACS, each muscle contributes to one or a number of AU(s), while a single AU can also be associated with more than one muscles. These muscles are related to each other dynamically and spatially, enabling a coherent and consistent facial expression (Ekman et al., 2002). Table 1 summarizes some AU examples, their associated facial muscles and corresponding emotions. The possible interpretations of emotions pertaining to each AU are also provided. By noticing specific changes of corresponding AUs, one can visually perceive and recognize each subtle facial expression.

2.2 Related Applications for Facial Emotion Detection

There has been extensive research focusing on automatic facial emotion recognition. Current approaches in the area can be categorized into two groups: *static* and *dynamic* feature based.

The static feature based systems usually focused on recognizing emotional facial expressions by observing representative facial geometric (e.g. points or shapes of facial components) or appearance features (e.g. facial wrinkles, furrows or bulges) statically and directly from the image data. For example, Soyel & Demirel (2007) extracted six characteristic distance features from the distribution of 11 facial feature points in a 3D facial model, and then employed them as inputs to a Neural Network classifier for the recognition of the six basic emotions. Rao et al. (2011) extracted grey pixel features from eye and mouth regions, and then used Auto-Associative Neural Network (AANN) models to capture the distribution of the extracted

features. Their system achieved an 87% average accuracy for emotion recognition from video inputs. Tang & Huang (2008) utilized 96 distance and slope features extracted from a cropped 3D face mesh model with 87 landmark points, and achieved an 87.1% average accuracy for the recognition of the six basic emotions by using multi-class Support Vector Machines (SVMs). Mahoor et al. (2011) employed Gabor coefficients transformed from 45 facial landmark points based on Active Appearance Model (Lucey et al., 2006), and classified AU combinations using a Sparse Representation (SR) classifier. Whitehill et al. (2011) detected 19 AUs by feeding 72 complex-valued Gabor filtered features to a separate linear SVM, and recognized six basic emotions using multivariate Logistic Regression (MLR) from the detected AUs. There are also some other facial action and emotion recognition approaches using static features that have been investigated, such as Local Binary Patterns (Shan et al., 2009) and Haar features (Whitehill & Omrin, 2006), etc.

The use of only static features, however, faces a drawback, i.e. the dynamic information of facial movements has been ignored and also the static features tend to vary a lot between different subjects (e.g. the shapes of eyes and the width of mouth). Thus it may lead to the inadequacy of generalization ability and efficiency. In order to address this issue, recently some research has made efforts in capturing dynamic facial features or making use of temporal variation of facial measurements. For example, Besinger et al. (2010) tracked 26 facial feature points from five facial image regions (eyebrows, eyes and mouth), and used displacements of them to recognize three basic emotions. Valstar et al. (2012) used Gabor-feature-based boosted classifiers and particle filtering with factorized likelihoods to track 20 facial points through a sequence of images. These facial geometric points were then used as inputs to a hybrid classifier composed of Gentle Boost, SVMs, and hidden Markov models (HMMs) to recognize 22 AUs. Wang & Lien (2009) employed 3D motion trajectories of 19 facial feature points as inputs to SVMs and HMMs for the recognition of seven AU combinations. Kotsia et al. (2008) recognized 17 AUs and seven emotions by the fusion of displacements of 104 Candidate grid nodes and texture information features using SVMs and Median Radial Basis Functions (MRBFs) Neural Networks. Tsalakanidou & Malassiotis (2010) proposed a rule-based automated AU and emotion recognition system based on facial geometric, appearance, and surface curvature features extracted from 2D+3D images. The results demonstrated good accuracy rates for the recognition of 11 selected AUs and four types of emotions. Srivastava & Roy (2009) used spatial displacements (or residues) of 3D facial points and SVM classifiers to recognize the six basic emotions, and demonstrated better recognition accuracies in comparison to the employment of pure static facial features (91.7% for dynamic features vs 78.3% for static features).

Although the above dynamic feature based systems showed noticeable improvements on recognition accuracy, and overcame some of the inherent defects in typical static feature based methods, many state-of-the-art AU and emotion recognition systems still suffered from the following difficulties. First of all, automatic AU intensity measurement posed great challenges to automated recognition systems since the differences between some AUs' intensity levels could be subtle and subjective, and

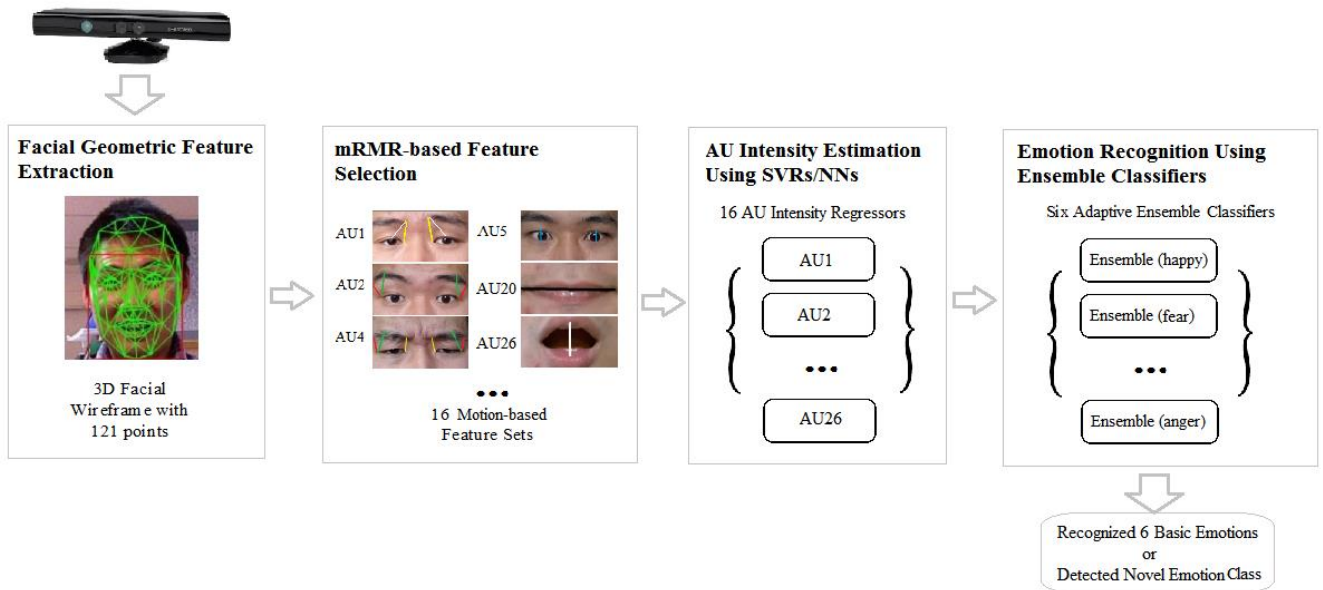


Figure 2. The overall system architecture and data processing pipeline

the physical cues of one AU might vary greatly when it occurs simultaneously with other AUs. Furthermore, FACS only defines a five point ordinal scale to describe the intensity of an AU. It does not define a quantifiable standard to measure the strength of corresponding facial changes. Hence, although there is substantial research concentrating on automatic AU recognition (e.g. Sorci & Thiran, 2010; Pantic & Patras, 2006; Tong et al., 2007; Li et al., 2013), the companion problem of accurately estimating the AU intensity levels has not been much investigated. There were only limited applications in the literature on AU intensity estimation. For instance, Kaltwan et al. (2012) realized continuous AU intensity estimation based on facial landmarks and appearance features by using a set of independent regression functions, but the work only focused on 11 specified AUs that are related to shoulder pain expressions. Bartlett et al. (2006) found that in AU classification tasks, distances between samples to SVM separating hyperplanes were correlated with AU intensities. Based on this finding, Savran et al. (2012) realized intensity estimation of 25 AUs from still images on both 2D and 3D modalities using appearance features and regression based methods. They claimed that the proposed approach for AU intensity estimation performed better than other state-of-the-art methods.

Furthermore, in contrast to AU detection, robust facial emotion recognition using AU intensities is still largely unexplored. Current approaches mainly focused on rule-based and statistical-based methods. For example, Valstar & Pantic (2006) explored both a formulated rule-based method and an Artificial Neural Network (ANN) based method to predict emotions from AUs. However, their recognition accuracies still required further improvements. It could be attributed to the fact that the former, i.e. the rule-based reasoning, was not robust enough to deal with noises and errors, while the latter, i.e. directly using machine learning techniques, relied on extensive training data to accommodate possible AU combinations for each emotion category. Chang et al. (2009) proposed a hidden conditional random fields (HCRFs) based method to map various combinations of 15 most frequently occurring AUs to underlying emotions, but extensive annotation work was required prior to map-

ping.

This paper aims to overcome these challenges discussed above, and develop a practical, robust and person-independent solution for facial Action Unit intensity estimation and emotion recognition. This research employs motion-based facial features with a strong psychological background to estimate the intensities of the 16 AUs closely associated with the expression of the six basic emotions. Subsequently, the 16 AUs are ranked for each emotion according to their discriminative power. The derived intensities of the most discriminative AU combinations are then employed as inputs to robustly recognize the six basic emotions regardless of errors and noises involved in the input AU intensities. The proposed system is discussed in detail in the following.

3 INTELLIGENT AU INTENSITY ESTIMATION AND FACIAL EMOTION RECOGNITION

In this section, we provide an overall description of the proposed facial emotion recognition system, which is composed of: facial geometric data tracking, mRMR-based feature selection, Action Unit intensity estimation using Neural Networks (NN) and Support Vector Regressors (SVR) and emotion recognition with ensemble classifiers. Figure 2 shows our system's overall architecture and dataflow.

1. The real-time facial geometric data tracking is implemented based on a Microsoft Kinect sensor (Webb & Ashley, 2012) and a variant of Candide-3 model (Ahlberg, 2001). The Kinect's facial analysis API is able to localize a total of 121 3D facial landmarks and perform continuous tracking at a frame rate of 25~30 fps.
2. We extract motion-based facial features for AU intensity estimation, which are calculated based on facial wireframe node displacements. We then apply both manual and mRMR based automatic feature selection methods to select 16 sets of informative features from the complete pool of candidate features for the 16 diagnostic AUs.
3. The feature sets selected by the mRMR based optimization are respectively employed as inputs to 16 AU intensi-

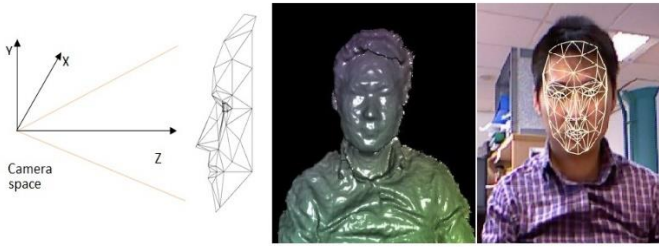


Figure 3. The Kinect 3D coordinate system (left), 3D surface reconstruction with depth data (middle) and a tracked 3D facial wireframe (right)

ty estimators, with each estimator dedicated to each AU. We employ Neural Networks and Support Vector Regressors for AU intensity estimation.

4. For robust emotion recognition, the 16 diagnostic AUs are first ranked and filtered according to the AU-Emotion relationships with intention to identify the most discriminative AU combinations for each emotion category. We then propose six novel adaptive ensemble models for robust classification of the six basic emotions and novel emotion detection, with each ensemble dedicated to each emotion category.

3.1 Facial Geometric Feature Tracking

Regarding to 3D facial geometric feature extraction, a number of well-known methods have been examined, such as the Kanade-Lucas-Tomasi (KLT) tracker (Bouquet, 1999) and the Vukadinovic-Pantic facial point detector (Vukadinovic & Pantic, 2005). Both of them can generate good tracking results with static input images, but limitations rise up when deal with real-time 3D streams. In our system, the 3D face geometric data are acquired through a Kinect and its embedded face tracking engine (Webb & Ashley, 2012). The Kinect is an effective research tool that physically integrates a color camera with up to 1280 x 960 resolutions, a depth-sensing camera with up to 640 x 480 resolutions, and an array of four microphones. It provides efficient real-time 3D tracking capabilities in a relatively inexpensive package.

When emotions are being expressed by a subject, the facial elements change their shapes and positions accordingly. These geometric changes caused by facial muscles contain rich motion-based facial features. Once completing parameter adjust-

ments and successfully detecting a user's face, the Kinect face tracking engine performs fitting and subsequently tracks a 3D variant of the Candide-3 model with 121 grid nodes. The facial tracking algorithm makes use of both color and depth image data streams to reconstruct salient facial models, enabling better robustness against variations in illumination, scaling, skin color and especially head poses. In good lighting conditions, it is able to track a face reliably when the user's head pitch, roll and yaw are respectively less than 10, 45 and 30 degrees (Webb & Ashley, 2012).

The tracked facial wireframe is able to automatically fit to the detected face in the Kinect 3D coordinate space and evolves through the video sequence (see Figure 3). It is able to reach up to 30 fps on i7 quad-core CPUs with 8GB RAM. If required, the loss or error of tracked wireframes could be handled by a model deformation algorithm, which is able to add mesh fitting at the intermediate steps of tracking. Such a procedure increases robustness against node losses and ensures tracking effectiveness. An essential normalization procedure is also performed afterwards, where the information of head orientation and distance to the sensor is employed to adjust the tracked facial grid model. Figure 4 shows a neutral state plus facial expressions for the six basic emotions associated with generated corresponding 3D facial wireframes.

3.2 Facial Action Unit Intensity Estimation

In literature, most recent research work employed either image driven or prior model-based methods for automatic AU recognition. The former (e.g. Chang et al., 2004) performed recognition based on static image data directly while the latter was developed to extract the relationships and spatial-temporal information of AUs using prior models (e.g. Tong et al., 2010; Valstar & Pantic, 2007). However both required a considerable amount of reliable training data, which sometimes could be difficult and expensive to acquire. More importantly, generalizing a model trained on one database to other databases could still be a challenging issue, especially for real-life applications (Li et al., 2013; Torralba & Efros, 2011). In order to overcome these challenges, we propose and employ motion-based facial features, which are supported by psychological studies and facial anatomy, and thus are more pertinent for AU intensity estimation. The 16 AUs we focus on in this research are AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller), AU13 (Cheek Puffer), AU15 (Lip Corner Depressor), AU17 (Chin Raiser), AU18 (Lip Pucker), AU20 (Lip Stretcher), AU23 (Lip Tightener), AU24 (Lip Pressor), AU26 (Jaw Drop) and AU27 (Mouth Stretch).

- We propose dynamic motion-based facial features (e.g. the elongation of mouth) for AU intensity estimation, which can be measured through the displacement of facial points between natural and expressive frames. As discussed earlier, such features are caused by underlying facial muscle movements, and thus are relatively universal and subject-independent.
- We apply both manual and automatic methods to select a unique subset of informative features for each AU respectively. The manual feature selection is guided by FACS domain knowledge, while the automatic feature selection

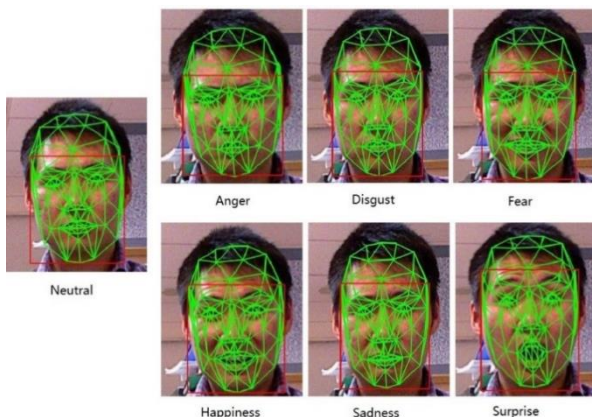


Figure 4. Examples of tracked 3D facial wireframes for each expression (The green lines represent facial wireframes, while the red rectangles indicate detected facial areas)

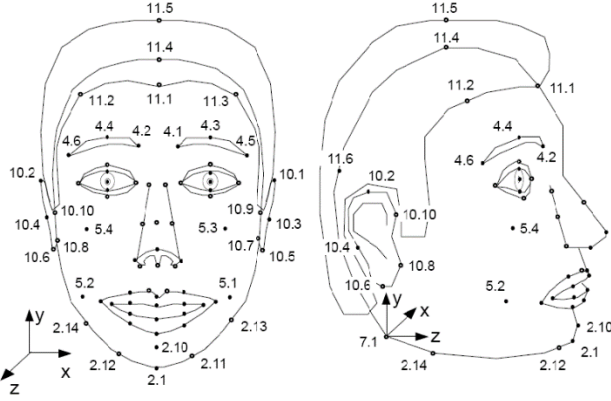


Figure 5. Facial feature points defined in MPEG-4 (Pandzic & Forchheimer, 2012)

is performed based on mRMR based optimization (Peng et al., 2005). Their performance and comparison are presented in Section 4.2.

3.2.1 Extraction of Motion-based Facial Features

As a part of MPEG-4 FBA [ISO14496] International Standard, the MPEG-4 face animation framework (Pandzic & Forchheimer, 2012) is designed to deal with face animation applications, including reproduction of facial shape, texture, subtle expressions, as well as speech pronunciation. MPEG-4 defines 84 facial feature points to best reflect the facial anatomy and movement mechanics, which are learned from subtle facial actions and are closely related to muscle actions, as illustrated in Figure 5 (Pandzic & Forchheimer, 2012). Based on this knowledge, we derive a series of 3D distance features between key facial points, and then use dynamic changes of these distances for AU intensity estimation.

When reliably detecting a user's face, the face tracking component continuously outputs a sequence of normalized 3D facial wireframes (compatible with MPEG-4 standard) in a real-world 3D coordinate system. Each wireframe consists of 121 grid nodes, including 16 nodes for eyes (i.e. 8 nodes for each eye contour), 20 nodes for eyebrows (i.e. 10 for each eyebrow), 12 nodes for the upper lip, 16 nodes for the lower lip, 16 for the nose, and others for making up the rest of the mesh model. The tracking process of 3D geometrical feature points is also robust to head rotations up to 10, 45 and 30 degrees in pitch, roll and yaw as discussed above.

We first acquire reference measurements of the neutral facial expression of each subject. Rather than requiring subjects to deliberately pose an initial calibration expression of the neutral state (which is often unreliable), we record the first 50-100 frames (typically 2-4 seconds, when subjects are naturally in their neutral states), and then compute the median data of these neutral frames to form a set of reference measurement vectors $\{R_i\}$ for the representation of neutral faces.

The motion-based facial features can be computed through facial point displacements between natural and expressive frames. Equations (1) and (2) define the calculation of any motion-based facial feature in the 3D Euclidean space.

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

$$\Delta d = d_{i,j}(\text{expressive}) - d_{i,j}(\text{neutral}) \quad (2)$$

In Equation (1), $d_{i,j}$ is the distance between $node_i$ (i.e. a 3D facial point i) and $node_j$ (i.e. a 3D facial point j) among the generated 121 3D facial wireframe nodes, and in Equation (2), Δd defines the change of distance feature $d_{i,j}$ between the reference (neutral) frame and any expressive frame. Such distance features are computed based on a real-world 3D coordinate system. As discussed earlier, the facial tracking engine of the Kinect is able to perform face fitting with high accuracy and is also able to identify the distances of different facial regions to the camera using depth images obtained from its depth camera to deal with facial point extraction with head rotations. Thus, our facial tracking component developed based on such a platform is capable of providing robust fitting and geometrical 3D feature extraction to deal with head pose variations and movements in real-life applications.

However, n number of facial feature points will result in a large number of C_n^2 unique distance features (e.g. 121 facial points will produce $C_{121}^2 = 7260$ distance features). Intuitively, not all of the distance features are informative for the detection of a specific AU. Thus, rather than applying the distance features between entire facial points for all AUs without distinction (e.g. Kotsia et al., 2008), we next step focus on generating a subset of informative discriminating features from the candidate feature pool for each AU respectively, which may lead to optimized performance.

3.2.2 Feature Selection for AU Intensity Estimation

Manual feature selection

In typical manual feature selection, the features are derived based on sufficient domain knowledge. We extract a total of 24 representative facial motion-based features (i.e. Δd distance changes) using 22 key facial feature points out of the whole 121 points, as illustrated in Table 2. According to Ekman & Friesen (1983) and Ekman et al. (2002), these features are believed to play an important role in determining the level of AU intensities. As shown in Table 2, each AU is associated with a subset of features composed of only a small number of relevant features (typically 2 to 6 dimensions). Such features are derived according to FACS domain knowledge, and we especially focus on analyzing the movement of facial muscles underlying each AU for subsequent AU intensity estimation.

Moreover, we provide two examples for manual feature selection in the following. For example, when AU1 (Inner Brow Raiser) is occurring for a specific facial emotion expression, the inner portion of the eyebrows is pulled upwards by muscle 1, see Figure 6 (Ekman et al., 2002). This causes an inevitable increase in the distance between inner eyebrow corner and inner eye corner. Thus, the distance variation Δd between the neutral

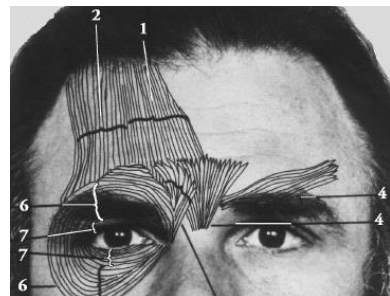


Figure 6. Muscles associated with upper facial Action Units (Ekman et al., 2002)

TABLE 2
EXAMPLES OF MANUALLY SELECTED FEATURES AND MEASUREMENTS REPRESENTED BY LINES OF DIFFERENT COLORS

AU	Measurement nodes	Distance Features (Neutral)	Distance Features (Expressive)
AU1 Inner Brow Raiser	Inner eyebrow corner		
AU2 Outer Brow Raiser	Outer eyebrow corner		
AU4 Brow Lowerer	Middle top of eyebrow		
AU5 Upper Lid Raiser	Middle eyelid top		
AU6 Cheek Raiser	Middle eyelid bottom		
AU10 Upper Lip Raiser	Right top of upper lip		
AU12 Lip Corner Puller	Outer eye corner		
AU15 Lip Corner Depressor	Mouth corners		
AU18 Lip Pucker	Right mouth corner		
AU20 Lip Stretcher	Right mouth corner		
AU23 Lip Tightener	Right/Left top of upper lip		
AU26 Jaw Drop	Middle bottom of upper lip		

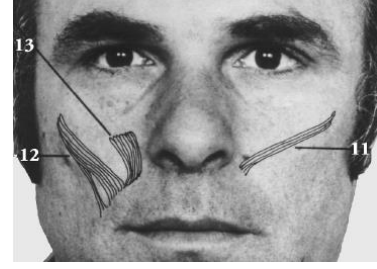


Figure 7. Locations of muscles underlying lower facial oblique Action Units (Ekman et al., 2002)

maticus Major (12) and *Caninus* (13), as shown in Figure 7 (Ekman et al., 2002). Both originate on the upper cheek bone and attach with the corner (angle) of the lips. When contracted, they will pull the corners of the mouth naturally up towards the upper cheek. Thus, the distances between mouth corners and outer eye corners are reduced synchronically. Therefore, we select eye corners as reference points because their positions are relative fixed and can be reliably tracked for AU12 or AU13.

Note that, in this research, Δd can be either positive or negative. For instance, AU1 (Inner Brow Raiser) may cause a positive Δd which means an increase in distance between inner eye corners and eyebrow corners. When Δd becomes negative, it indicates the eyebrow is lowered, which means AU4 (Brow Lowerer) occurs. Table 2 summarizes some AUs and their corresponding manually selected features, and gives a clear illustration on how they change synchronically with the occurrence of each AU (for clarity, all samples showed in Table 2 are in 2D although in the real system, 3D facial points are extracted as discussed in Section 3.1). The above FACS domain knowledge-based manual feature selection provides an efficient and robust approach against facial shape variations of different subjects.

Automatic feature selection based on mRMR







Although equipped with appropriate domain knowledge, manual feature selection is often time consuming and requires endless trial-and-error process. There are also extensive optimization algorithms and boosting techniques devoted to automatic feature selection and feature dimensionality including Principle Component Analysis (PCA), Fisher Linear Discriminant (FLD), genetic and evolutionary algorithms, and AdaBoost etc. PCA has been widely used for feature selection for face and facial expression recognition for decades (Jong et al., 2009). According to Swets & Weng (1996), PCA derives most expressive features but may not embed sufficient discriminating power. FLD is another commonly used feature reduction technique which is claimed to provide comparatively more class separability by maximizing the mean between classes and minimizing the variation within a class (Chavan & Kulkarni, 2013; Gu et al., 2012). However, it requires a wide coverage of face/class variations at the training stage in order to get more superior recognition performance.

As the most common form of evolutionary optimization, conventional genetic algorithms evolve a large population of candidate solutions by mimicking the process of natural selection (Sikora & Piramuthu, 2007). Other commonly used evolutionary algorithms include Particle Swarm Optimization (Wang et al., 2007) and Genetic Programming (Davis et al., 2006), etc. However, applying such algorithms in a large search space (e.g. thousands of dimensions) may tend to be very computationally

and this expressive frame may contribute to the estimation of the occurrence and intensity of AU1.

Furthermore, the following indicates a slightly more complicated example. AU12 (Lip Corner Puller) and AU13 (Sharp Lip Puller) are often accompanied by a smile or a joyful facial expression. These AUs are caused by pulling the corners of the lips back and upwards to form a \smile shape of the mouth. But it is unlikely that we can directly use some intuitive distance features, such as the elongation of the mouth, to distinguish these AUs (although the mouth is indeed elongated). The reason is that there are other AUs that can also cause mouth elongation, such as AU20 (Lip Stretcher). Thus the extraction of distance features becomes challenging. However by analyzing these facial movements from the perspective of anatomy, we can see there are two underlying muscles related to these AUs - *Zygo-*

TABLE 3
COMPARISON OF MANUALLY SELECTED FEATURES WITH THOSE AUTOMATICALLY SELECTED BY mRMR

	<i>Manually Selected Features</i>	<i>Automatically Selected Features</i>
<i>AU1 Inner Brow Raiser</i>		
<i>AU2 Outer Brow Raiser</i>		
<i>AU18 Lip Corner Puller</i>		

exhaustive and time consuming. Furthermore, inappropriate parameter configuration may easily lead to premature convergence to a local extremum. On the contrary, mutual information (MI) is information based feature selection that is not limited to linear dependencies, and is able to maximize information in a class. Research on the performance improvement of MI has brought to the development of minimal-redundancy-maximal-relevance criterion (i.e. mRMR), which is a variant of MI. In this research, since a large proportion of the raw facial distance features could be less informative or considerably redundant with each other, it is reasonable to apply information theory based methods for automatic feature selection, which could well reflect relevance between features and outputs and within features comprehensively. Moreover, such methods also have relatively lower computational complexity and better generalization of the selected features on different classifiers. Thus, we are motivated by mRMR to propose an attractive alternative for automatic feature selection.

Moreover, Tang and Huang (2008) proposed a novel method based on maximizing the average relative entropy of marginalized class-conditional feature distributions, and successfully applied it to 3D facial distance feature selection tasks. Their automatically selected features achieved higher recognition accuracies than their manually devised features for the six basic emotions (about 2% - 5% improvements). However, their method is difficult to be applied to regression problems as the lack of effective relevant calculation method for continuous values. Thus, we introduce a modified mRMR-based feature selection method to deal with the case where both features and outputs are continuous data.

We introduce the mRMR optimization algorithm in the following. mRMR is introduced by Peng et al. (2005) and aims to minimize the mutual information between the selected features (i.e. redundancy), and to maximize the mutual information between the selected features and the desired output (i.e. relevance). Let x_i denote a feature and $S_M = \{x_i\}_{i=1}^M$ be an instance consisting of M features. I denotes the mutual information with y indicating the desired output, and $p(x_i)$, $p(y)$, $p(x_i, x_j)$, and $p(x_i, y)$ representing the probabilistic density functions. Then the traditional mRMR measure can be described as follows:

$$mRMR(i) = I(x_i, y) - \frac{1}{M-1} \sum_{x_j \in S_M, j \neq i} I(x_i, x_j) \quad (3)$$

where

$$I(x_i, y) = \sum_{x_i \in X_i} \sum_{y \in Y} p(x_i, y) \log \left(\frac{p(x_i, y)}{p(x_i)p(y)} \right) \quad (4)$$

Since both the features and AU intensities in our system are continuous values, their mutual information is often hard to compute. I.e. it is difficult to compute the integral in the continuous space using a relatively limited number of samples. One solution is to perform a uniform data discretization processing in advance of the estimation of the mutual information value. However, this may lead to considerable information loss.

An alternative solution is to use linear correlations to approximate the mutual information, as suggested by Metallinou et al. (2013). Here, by replacing the traditional mutual information metric with the Pearson correlation coefficient (CORR), the mRMR measure can be well adapted to continuous values. The CORR represents the linear relationship between a pair of values, defined as follows:

$$CORR(x, y) = \frac{COV\{x, y\}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5)$$

where COV stands for the covariance, and σ stands for the standard deviation, while $\bar{}$ symbolizes the mean.

Specifically, let $CORR(x_i, x_j)$ and $CORR(x_i, y)$ denote the linear correlations between two selected features, x_i , x_j , and between a feature x_i and the desired output y , respectively. The linear correlation based mRMR measure can be defined as follows:

$$mRMR(i) = |CORR(x_i, y)| - \frac{1}{M-1} \sum_{x_j \in S_M, j \neq i} |CORR(x_i, x_j)| \quad (6)$$

Then we perform a ranking of features according to their mRMR values. A higher value is preferred and it indicates that a specific feature contains more discriminating information, i.e. it has higher correlation with the desired output and lower correlation with other features. We try different numbers of top ranking features as the inputs for AU intensity estimation, and those leading to the best performance are determined as the optimal features for each AU regression, respectively. Table 3 illustrates some examples of the automatically selected features. Evaluation results indicate that the proposed mRMR-based feature selection yields comparable results for AU intensity estimation when compared with the manual feature selection process.

3.2.3 AU Intensity Estimation using Motion-based Features

For the construction of automatic AU intensity estimation, we notice the following challenges. First, because of individual differences among subjects, overlapping between intensity levels (Savran et al., 2012) and annotators' subjectivity are inevitable. Second, the relationship between AU intensity levels and the scale of evidence might be nonlinear. To solve these problems, we employ two widely accepted algorithms, feedforward Neural Network with Backpropagation (Hecht-Nielsen, 1989) and Support Vector Regression (Vapnik, 1995) for AU intensity estimation, because of their effective handling of data comprising noises and non-linear relations. We also aim to examine the effectiveness of the mRMR based optimization in comparison to the manual feature selection, and to determine whether the features selected by mRMR are effective enough for discriminating between different levels of AU intensities.

ALGORITHM 1
THE TRAINING ALGORITHM OF THE NEURAL NETWORKS FOR AU
INTENSITY ESTIMATION

- Create a feed-forward network with i input units, h hidden layer units, and one output unit o .
- Set all unit weights w_u using initial random values (e.g. decimals between -1 to +1).
- Set a proper small learning rate value r , ranging from 0 to 1 (e.g. 0.1).
- Until the termination condition (error < a set threshold value or reaching the number of the maximum iterations) is fulfilled, do.
 For each training dataset, do.

Propagate the input forward through the network:

- 1) Input each Δd_i to the network and compute the output o_u of every unit u of the network.

Propagate the errors backward through the network:

- 2) For the network output unit o , calculate its error e_o .

$$e_o = (I - o) * g'(o)$$
- 3) For each hidden unit h , calculate their error values e_h .

$$e_h = g'(o_h) * \sum_{i \in \text{outputs}} (w_{i,h} * e_o)$$

where g' is the first derivative of the sigmoid function

- 4) Update each network weight $w_{j,k}$

$$w_{j,k} = w_{j,k} + \Delta w_{j,k}$$

$$\Delta w_{j,k} = r * e_j * x_{j,k}$$

Feedforward Neural Networks for Regression

A feedforward Backpropagation Neural Network (BPNN) has the following two characteristics well suitable to our application:

- It is robust to the noise and errors involved in training data, which may be inevitable in many supervised applications as mentioned above (Mitchell & Hill, 1997).
- It needs some training costs, which depend heavily on the sample size, the dimensions of the training data, and the accuracy requirements. Once the model trained, however, it is extremely fast to be applied to the subsequent test instances. This would be beneficial to our real-time application.

A continuous value ranging from 0 to 1 is used as the single output to cover the whole interval of AU intensity levels ('0' represents absence with '1' indicating maximum AU intensity).

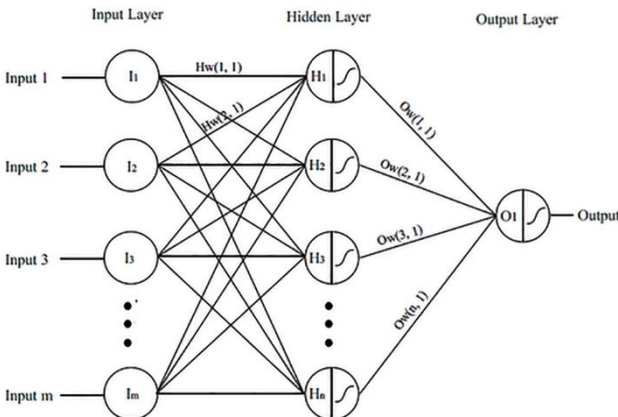


Figure 8. A typical topology of a feedforward Neural Network (Hecht-Nielsen, 1989)

In this way, we can preserve sufficient AU intensity information for subsequent emotion recognition. Thus, we have the training data format as follows:

$$dataset_n = \{\Delta d_1, \Delta d_2, \Delta d_3, \dots, \Delta d_i, I\}$$

where the inputs Δd are the informative motion-based facial features for each AU selected by either the manual process or the mRMR-based optimization, and the output, I , is the ground truth intensity of that AU. Both the training and testing datasets are scaled using the same procedure before applied into Neural Networks in order to achieve the best performance (i.e. linearly scaling each attribute to the range of [-1; +1] or [0; 1]).

We implement 16 three-layer feedforward Neural Networks. Each of them has an input layer, a hidden layer with 3 - 6 nodes based on the complexity of the input layer, and an output layer. We also adjust the learning rate, the momentum and the termination error parameters to modest values (e.g. respectively 0.1, 0.8, and 0.01) to best achieve a balance between accuracy, speed and generalization performance. Figure 8 illustrates an example topology of the applied feedforward Neural Network. Algorithm 1 lists the learning mechanism of the Backpropagation algorithm.

Support Vector Machines for Regression

Support Vector Machine (SVM) is a powerful machine learning algorithm based on minimizing the generalization error bound (structural risk) rather than minimizing the observed training error (empirical risk), so as to achieve better performance. The basic idea of Support Vector Regression (SVR) is to compute a linear regression function in a higher dimensional feature space where the lower dimensional input data are mapped using a kernel function (Basak et al., 2007).

Given training dataset as:

$$\{(x_1, y_1), \dots, (x_\ell, y_\ell)\} \subset \mathcal{X} \times \mathcal{R}$$

where x_i and y_i indicate the attribute and target values respectively, and \mathcal{X} denotes the space of the input patterns (e.g. $\mathcal{X} = \mathcal{R}^d$). In epsilon-SVR, the goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time as flat as possible. In simple linear case, $f(x)$ has the form as:

$$f(x) = \langle w, x \rangle + b \text{ with } w \in \mathcal{X}, b \in \mathcal{R} \quad (7)$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product in \mathcal{X} , and b indicates a bias value. *Flatness* in (7) means seeking a small vector w . To ensure this, one way is to minimize the Euclidean norm i.e. $\|w\|^2 = \langle w, w \rangle$. By introducing slack variables ξ_i, ξ_i^* to cope with infeasible constraints in some practical cases or allow for some errors, this problem can be written as the following formulation (8):

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \\ & \text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \quad (8)$$

where ξ_i, ξ_i^* denote the allowed upper and lower error bound respectively and the constant $C > 0$ determines the tradeoff between the flatness of f and the amount up to which deviations larger than ε are tolerated. This corresponds to dealing with the ε -intensive loss function described by (9) (Vapnik, 2001):

$$|\xi|_\varepsilon := \begin{cases} 0 & \text{if } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{otherwise.} \end{cases} \quad (9)$$

By constructing a Lagrange function and utilizing Lagrange multipliers, the original problem can be solved. The objective function can be rewritten as follows (Vapnik, 2001):

$$f(x) = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b. \quad (10)$$

where α_i, α_i^* are computed Lagrange multipliers. Here, by using a nonlinear kernel function $k(x_i, x)$ satisfying Mercer's condition instead of the dot product $\langle x_i, x \rangle$ in (10), SVR can be employed for nonlinear regression.

As advances in statistical learning theory, Support Vector Regression shows two capabilities that well meet our requirements:

1. SVR is especially suitable for the regression problems for a small sample size. The establishment of facial databases, especially the manual annotation is an expensive process, therefore it is necessary to maximize the use of limited amount of data.
2. The structural risk minimization principle endows SVR with good generalization capability for unseen data, thus the robustness and adaptation to different subjects of the system are enhanced.

We employ the established LibSVM Library (Chang & Lin, 2011) for the SVR implementation. We apply 16 epsilon-SVRs for the regression of the 16 selected AUs respectively, using the same input/output data format as discussed above. A scaling procedure is also performed before applying SVRs to achieve the best performance.

Kernel selection plays a key role for SVR model, since using different kernels may significantly influence the performance when dealing with the same problem. For this research, we consider the non-linear radial basis function (RBF) kernel as a reasonable choice, because:

1. RBF nonlinearly maps inputs into a higher dimensional space, thus it can well handle the case that the relation between facial features and AU intensity levels is nonlinear.
2. RBF has fewer number of hyperparameters than other nonlinear kernels (e.g. polynomial kernel), which may reduce the complexity of model selection (Hsu et al., 2010).
3. RBF usually has lower computational complexity, which in turn indicates better real-time computational performance.

Please note that when the dimensions of features are very high (e.g. thousands), the RBF kernel may become not suitable in comparison to a linear kernel (Hsu et al., 2010). However, it is not the case in this application.

Once the RBF kernel is selected, an essential step is to find optimized sets of cost (C), gamma (g) and epsilon (ε) parameters. We perform a "grid search" procedure on those parameters using the cross-validation technique, since it is regarded as one of the most effective methods to prevent over-fitting. In ν -fold cross-validation, the overall dataset is firstly divided into ν groups with equal number of samples in each group, then we use $\nu-1$ groups of the data for training and the remaining group for testing. This process is repeated ν times so that each group can be tested in turn. Specifically, various combinations of parameter values (i.e. exponentially growing values: $C = 2^{-10}, 2^{-9}, \dots, 2^{15}$; $g = 2^{-15}, 2^{-14}, \dots, 2^{10}$; $\varepsilon = 2^{-10}, 2^{-9}, \dots, 2^{-1}$) are conducted

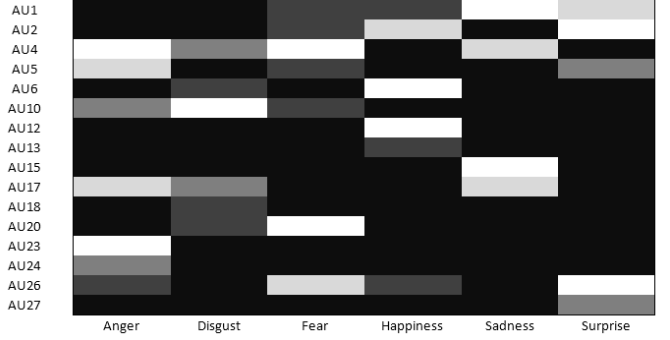


Figure 9. The AU-Emotion relation confusion matrix (lighter color indicating higher Influence Power with darker color representing lower Influence Power)

and the one with the lowest Mean Squared Error (MSE) under 5-fold cross-validation is selected. The MSE evaluates the prediction results by taking into account the squared error of the predicted value from the ground truth and can be computed as follows (DeGroot & Schervish, 2011):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2 \quad (11)$$

where y_i is the predicted value, and \tilde{y}_i is the ground truth. Moreover, the Pearson correlation coefficient is also employed to evaluate the linear relationship between the prediction and the ground truth, i.e. how they change together.

Thus, 16 Neural Networks and 16 SVRs are developed to estimate the intensity for each AU respectively. Both manually and automatic selected features are used as inputs respectively to NNs and SVRs to measure the intensities of 16 AUs. 729 3D facial scans extracted from the Bosphorus database (Savran et al., 2008) from 56 subjects are used for performance evaluation. The databases, experiments and evaluations are detailed in Section 4.

3.3 Facial Emotion Recognition using AU intensities

The mapping between AU intensities and emotions could be a challenging task. For example, a 'surprised' facial expression may indicate the presence of {AU1, AU2, AU5, AU26}, or the physical cues of {AU1, AU2, AU26} in different cases. The intensities of these present AUs could be also variable. These practical issues make deterministic rule-based techniques less effective (e.g. using translating formula: surprise = AU1+AU2+AU5+AU26 (Ekman et al., 2002)). Likewise, directly applying machine learning algorithms could be still very challenging, since extensive training data are needed to accommodate various possible combinations of AUs for emotional expressions. There are, however, more than thousands of possible AU combinations in spontaneous facial expressions (Ekman & Friesen, 1983), which are far beyond the data available in any existing databases. In order to deal with such challenges, we propose a novel method to robustly map AU intensities to the six basic emotions using a limited number of samples, which consists of two steps: (1) AU-Emotion relationship mining and ranking; (2) facial expression recognition using the identified discriminative AU combinations.

3.3.1 Mining and Ranking AU-Emotion Relationships

Rather than using the full set of 16 AUs for emotion interpretation indiscriminately, we first derive AU-Emotion relationship, and then identify the most effective combination of AUs as the

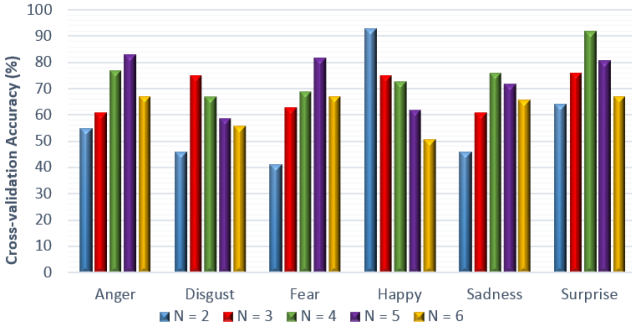


Figure 10. Average classification accuracies using SVMs and top ranking AUs ($N = \{2, 3, 4, 5, 6\}$) as inputs for the six basic emotions

discriminative AU set for each emotion category for subsequent emotion recognition. The AU-Emotion relationship is derived through statistical analysis of sufficient amount of valid samples with AU intensity and emotion annotations provided by the extended Cohn Kanade (CK+) (Lucey et al., 2010) and Bosphorus databases (Savran et al., 2008).

A new concept, Influence Power, is proposed to describe the weights of the AU-Emotion relationship, as defined in Equation (12):

$$P = (\sum_{i=0}^n Intensity_{x,i})/n \quad (12)$$

where n is the number of examples belonging to a given emotion category, $Intensity_x$ donates the intensity value of AU x occurred corresponding to the given emotion, and the magnitude of P quantifies the Influence Power of AU x for that emotion category. A higher Influence Power represents a closer connection between an AU and an emotion, while a lower value may indicate the weak association between them. 1200 samples (equally distributed to the six basic emotions) collected from the CK+ (Lucey et al., 2010) and Bosphorus databases (Savran et al., 2008) have been taken into account for AU-Emotion relationship identification. After normalizing P across all of the 16 AUs for each emotion, we draw the relation confusion matrix between the 16 AUs and the six basic emotions in Figure 9. Thus, a set of association weights between AUs and emotions is established.

Having obtained the relation confusion matrix, we then select the top N AUs with the highest Influence Power for the recognition of each emotion. On the positive aspect, this may significantly reduce the potential negative impact of those non-dominant or haphazard AUs and improve classification accuracy. For example, ‘happy’ expressions have AU6, AU12 as highly weighted associations with AU2 as a comparatively lower weighted association, while AU2 is also served as a key physical cue and thus has a higher association weight for ‘surprise’

TABLE 4
IDENTIFIED DISCRIMINATIVE AU SETS FOR THE SIX EMOTIONS

Emotions	discriminative AU Sets				
Anger	AU 4	AU 5	AU 17	AU 23	AU 24
Disgust	AU 4	AU 10	AU 17		
Fear	AU 1	AU 4	AU 10	AU 20	AU 26
Happy	AU 6	AU 12			
Sadness	AU 1	AU 4	AU 15	AU 17	
Surprise	AU 1	AU 2	AU 26	AU 27	

expressions. However, on the negative aspect, over-filtering those AUs with lower Influence Power may also increase the risk of information loss. Thus, in order to optimize the selection of the N number of AUs, we perform a series of experiments with different N number of AUs (i.e. using different numbers of top ranking AUs as inputs) for each emotion category. The AU combinations with the best recognition accuracy will be finalized for subsequent emotion classification. The details are discussed in the following.

3.3.2 Selection of the Most Discriminative AU Combination for Each Emotion

We employ six SVM classifiers for the recognition of the six basic emotions, with each classifier dedicated to one emotion category and employing a unique set of discriminative AUs as inputs. The selection of the discriminative AU combinations is detailed as follows:

We first perform emotion recognition using different numbers of top ranking AUs (i.e. $N = \{2, 3, 4, 5, 6\}$) as inputs, and record the recognition accuracies in each round. Specifically, for each classifier, we collect 120 samples in total, 50 from the CK+ database (Lucey et al., 2010) and 70 from the Bosphorus database (Savran et al., 2008), covering both positive and negative cases (presence/absence of that emotion) with roughly equal quantities. We also apply a 5-fold cross-validation scheme depending on the sample size. The average cross-validation accuracies obtained by SVM classifiers are summarized in Figure 10 (the other classifiers yield very similar patterns, thus are omitted in the Figure).

Based on the results shown in Figure 10, the AU combination leading to the best recognition accuracy is determined as the most discriminative AU combination for each emotion. These AU combinations are summarized in Table 4 and employed respectively as the finalized inputs for the six emotion classifiers. For example, in Figure 10, since the highest recognition accuracy for ‘anger’ is achieved when N equals to 5, we select the top five ranking AUs as the discriminative AU combination, i.e. AU4, AU5, AU17, AU23 and AU24. Thus, the derived intensities of these five AUs are subsequently used as inputs to the ‘anger’ emotion classifier. The discriminative AU combinations for other emotion categories are also determined as above. The experimental results and evaluations are presented in Section 4.

3.3.3 Emotion recognition using adaptive ensemble classifiers

In this research, we propose an adaptive ensemble scheme for the detection of six expressions and any newly arrived novel emotion classes. In this scheme, there are six ensemble classifi-

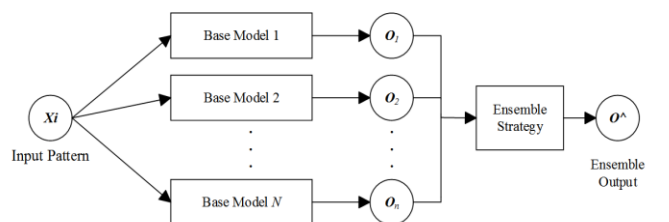


Figure 11. An example of an ensemble learning model

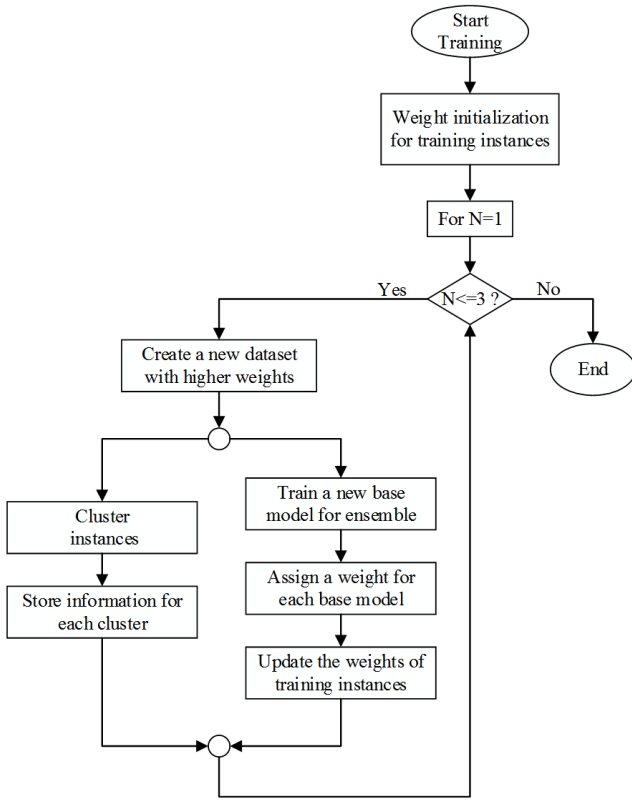


Figure 12. Flow chart of the generation of the proposed ensemble model

ers with each ensemble robustly differentiating the presence/absence of each emotion. We also employ single Support Vector Machines (C-SVC) classifiers to conduct the same expression recognition tasks, and their results will be used as the benchmark for comparison with those achieved by the ensemble classifiers.

Ensemble learning generally refers to approaches that generate several base models that are combined to make a prediction, as illustrated in Figure 11. Compared to traditional single model-based methods, ensembles have the advantages of improved robustness and increased accuracy (Garcia-Pedrajas et al., 2005). For an exhaustive review of ensemble approaches, readers may refer to Rokach (2010).

In the field of facial emotion recognition, many ensemble approaches have been proposed. For example, Whitehill & Omlin (2006) employed AdaBoost algorithm for AU recognition using Haar features. More recently, Zavaschi et al. (2013) created a pool of base SVM classifiers with features extraction conducted by Gabor filters and Local Binary Patterns, and then applied a multi-objective genetic algorithm to find the best ensemble by minimizing both the error rate and the size of the ensemble. Although ensemble models have been used for facial expression recognition, few of them are developed to detect novel emotion classes. Moreover, in the field of data stream mining, most of the existing ensemble algorithms integrated with novel class detection employed classic decision tree (e.g. Farid et al., 2013) or k-nearest neighbor (e.g. Masud et al., 2011) classifiers as their base models. In our research, we employ a special type of Neural Network, i.e. Complementary Neural Network, as the base classifier and propose a novel mechanism to further improve the performance of the 6-class emotion recognition and

novel emotion detection. The details of our approach are discussed as follows.

Each of the proposed ensemble classification models consists of two phases: ensemble model generation (training) and classification with novel class detection (testing). Figure 12 illustrates the work flow of the generation of an ensemble classifier. It starts with the weight initialization procedure for each training instance based on the posterior probability, as detailed in Section 3.3.3.1. Afterwards, the ensemble model generates a new training subset from the original training set using instances with higher weights. Then, a base model is trained using the newly generated training subset. Here, we employ a novel Complementary Neural Network (CMTNN) as the base classifier, because of its ability to estimate the vagueness level of classification results. The CMTNN is introduced in Section 3.3.3.2. A weight is subsequently calculated and assigned to the current base CMTNN classifier based on its classification accuracy rate for the original training dataset. We also update the weights of the original training instances with the goal of increasing the weights of those misclassified instances. The weight calculation and update methods are discussed in Section 3.3.3.3. The generated training subset is also clustered based on the similarities and differences of the instances, as discussed in Section 3.3.3.4. We employ the following idea for novel emotion class detection. A distance-based clustering technique and the vagueness measure of the classification results obtained by CMTNN will be employed to identify the arrival of novel emotion class (i.e. unseen expressions absent from the training set). Overall, the above procedures iterate three times, thus three weighted base models are generated (considering a balance between performance and computational complexity). The final ensemble classification results can be obtained by using majority of weighted votes of the three base models.

Moreover, Figure 13 shows the flow chart of classification and novel emotion class detection. As mentioned above, the proposed ensemble scheme is expected to effectively detect novel emotional expressions. Such capability is achieved by the analysis of both the vagueness values of the based models and the corresponding similarity-based clustering results. More specifically, once a testing instance arrives, the three base models for each ensemble respectively output both the individual classification results and the vagueness/uncertainty estimation values of the results. If any of the three vagueness values is greater than a threshold and the instance does not belong to any existing data clusters, then the instance is identified as a potential novel emotion class and will be stored in a separate dataset. Finally, if this instance is identified as a potential novel emotion by more than half of the ensemble classifiers of the six basic emotions (e.g. more than three ensembles), then it is determined as a newly arrived novel emotion.

3.3.3.1 Weight Initialization for Training Instances

First of all, we present the method on how to initialize the weight of each training instance based on naïve Bayes (NB) classifier. Although traditional ensemble approaches (e.g. boosting algorithms) normally initialize the weight of each training instance with an equal value, assigning appropriate weights using non-equal values has been also proved to improve the performance of ensemble classifiers (e.g. Farid et al., 2013).

In this research, the weight of each training instance is initial-

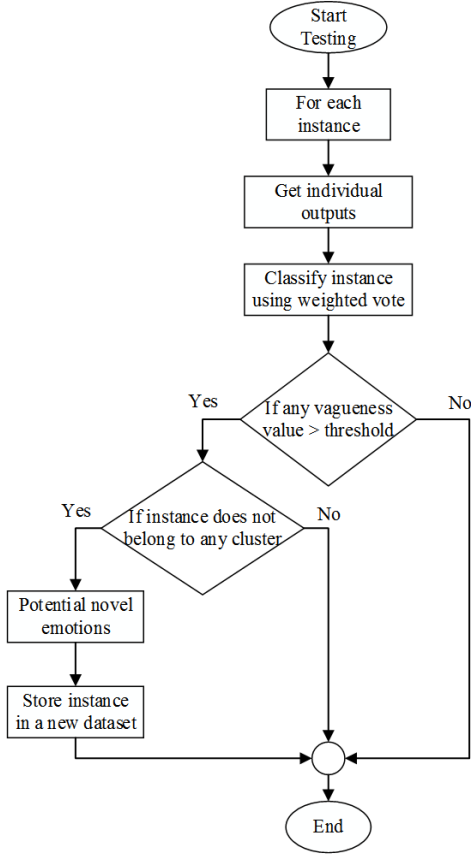


Figure 13. Flow chart of classification with novel emotion detection

ized based on the posterior probability obtained by a NB classifier. Specifically, we first estimate the prior probability $P(C_i)$ for each class C_i , by calculating how often each class occurs in the given training dataset. Similarly, for each attribute A_j and each class C_i , the class conditional probability $P(A_j|C_i)$ can be obtained by counting how often each attribute value occurs in each class. Given an instance x_i , assuming all attributes are independent, the conditional probability $P(x_i|C_i)$ can be estimated by combining the effects of each different attribute as shown in the following equation:

$$P(x_i|C_i) = \prod_{j=1}^n P(A_j|C_i) \quad (13)$$

Then, the posterior probability $P(C_i|x_i)$ can be calculated according to Bayes' theorem as:

$$P(C_i|x_i) = \frac{P(x_i|C_i) P(C_i)}{P(x_i)} \quad (14)$$

Thus, the posterior probability is obtained for each class. We then assign a weight for the instance x_i using the highest posterior probability. The weights of the rest instances are initialized using the same method. Once the weights of all instances are initialized, their weights will be normalized so that their sum equals to 1.

3.3.3.2 Base Model Generation (CMTNN)

Having initialized the weight for each training instance, we focus on the generation of each base model. Here, we introduce a Complementary Neural Network (CMTNN) as the base classifier, which is not only especially suitable for binary classifica-

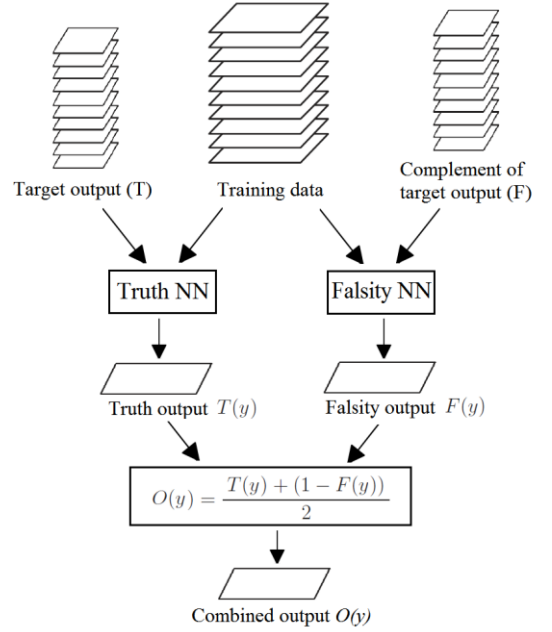


Figure 14. Topology of a Complementary Neural Network (Kraipeerapun, 2008)

tion problems, but also able to provide vagueness estimation of the classification results.

CMTNN, originally proposed by Kraipeerapun (2008), consists of a pair of opposite feedforward Neural Networks with the same architecture (i.e. a truth Neural Network and a falsity Neural Network). The truth Neural Network is trained by original training data to predict the degree of the truth membership values, and the falsity Neural Network is trained to predict the degree of the false membership values using the same inputs but the complement of target outputs of the original training instances (as illustrated in Figure 14). For instance, if the target output of original training data is 1, the complement of this target output used to train the falsity Neural Network should be 0.

For each test pattern, a CMTNN outputs both the truth and false membership values, and they are supposed to be complementary to each other ideally (i.e. if the truth membership value is 1 then the false one is supposed to be 0, or vice-versa). In practice, however, both membership values predicted may not always be informative enough for the final classification. For example, both the truth and false membership values are around 0.5. Thus, an uncertain classification occurs. Empirically, the greater proximity of the truth and false membership values, the higher the degree of vagueness exists. Given a testing pattern, let y_i be the output. $T(y_i)$ denotes the truth membership output, and $F(y_i)$ denotes the false membership output, then the vagueness value of the prediction $V(y_i)$ can be estimated as:

$$V(y_i) = 1 - |T(y_i) - F(y_i)| \quad (15)$$

By combining $T(y_i)$ and the complement of $F(y_i)$ using a simple equal weighted method, the final output $O(y_i)$ for the pattern can be calculated as:

$$O(y_i) = \frac{T(y_i) + (1 - F(y_i))}{2} \quad (16)$$

A threshold value is applied to Equation (16) to classify the

output into binary classes (generally, the most common threshold value is 0.5). An output pattern is classified as 1 (true) if $O(y_i)$ is greater than the threshold value, otherwise, it is classified as 0 (false). Compared to other traditional methods which solely apply truth membership values, CMTNN has two outstanding features: improved classification accuracy for binary problems and the ability to assess uncertainty of classification using the vagueness value (Jeatrakul & Wong, 2009).

3.3.3.3 Weight Calculation and Update

We then introduce the weight calculation methods for both of the base classifiers and training instances. First, once a base classifier is generated, a weight will be assigned based on its classification accuracy rate for the original training instances. Once all the three classifiers are generated, their weights will be normalized so that their sum equals to 1.

Moreover, for training instances, we follow the following steps to update their weights, with the intention to increase the weights of those instances which are more difficult to classify (i.e. those with higher error rates). We first assign an error rate for each training instances x_i by

$$error(x_i) = \begin{cases} 1, & \text{if misclassified} \\ 0, & \text{if correctly classified} \end{cases} \quad (17)$$

We then calculate the overall error rate for all instances as follows:

$$error_{overall} = \sum_{i=1}^n w_i * error(x_i) \quad (18)$$

where w_i is the current weight for instance x_i . Afterwards, the weights of the correctly classified instances will be decreased as follows:

$$w_{i,updated} = w_i * \left(\frac{error_{overall}}{1 - error_{overall}} \right) \quad (19)$$

Thus, the weights of correctly classified instances are decreased and the weights of those misclassified ones become increased comparatively. Once the weights of all instances are updated, their weights will be normalized, so that their sum remains the same as it was before.

3.3.3.4 Distance-Based Data Clustering

Clustering is a widely-used unsupervised learning technique. It is a main task of exploratory data mining, and has been applied to many application domains such as image analysis, pattern recognition, information retrieval, medicine, and bioinformatics. It is a form of learning by observation, and aims to determine the intrinsic grouping for a set of unlabeled data based on the principle that instances in the same group (called a cluster) are similar (or related) to each other and different from (or unrelated to) the instances in other groups. The greater the difference between clusters, and the greater the similarity within a cluster, the better the clustering.

In the distance-based clustering, we use the Euclidean distance as the metric to determining the similarity (or differences) of two instances. For a given instance x_i , if we can find any instance x_j in an existing cluster N that fulfills: 1. the Euclidean distance $D_{i,j}$ between x_i and x_j is minimum, and 2. $D_{i,j} < \alpha$ a predetermined threshold, the instance x_i is assigned to N . Otherwise, x_i is assigned to a newly generated or any other cluster. During the training phase, the distance-based clustering is em-

ployed to specially measure the distribution of the training instances. During the testing phase, if the output uncertainty level (i.e. the vagueness value of a CMTNN) of an instance is greater than a predetermined threshold, this instance will be further determined by the distance-based clustering. If the instance does not belong to any existing clusters, it is confirmed as a potential novel class.

4 EVALUATIONS AND DISCUSSION

In this section, we perform two types of evaluations of the proposed system: static off-line and real-time on-line evaluations. The off-line evaluation is purely based on annotated facial images borrowed from the Bosphorus database, for which we conduct exhaustive experiments for both AU intensity estimation and emotion classification to evaluate the system performance. The on-line testing mainly focuses on the assessment of the system's real-time performance and newly arrived novel emotion class detection, where we use the system trained with the database images to recognize facial expressions of real human subjects in real time.

4.1 Databases

In this research, we employ two facial expression image databases. The first database employed is the CK+ database (Lucey et al., 2010), which is based on 2D facial images but provides rich AU intensity and expression annotations. However, this database is only used for the statistical computation of the discriminative AU sets for each emotion as discussed in Section 3.3. The second database employed for this research is the Bosphorus 3D Database (Savran et al., 2008), which contains both 3D facial scans and manually labeled landmarks, as well as a large variety of Action Unit and expression annotation. This database is used for the evaluation of both AU regression and emotion classification. The introduction of these two databases is provided in the following.

1. **The Extended Cohn-Kanade Database** consists of 593 image sequences across 123 subjects with each image sequence starting from a neutral expression and ending in a peak frame emotional expression. Among 593 image sequences, the annotations of the six basic emotions and facial AUs are provided for 327 peak frame images. The AU annotations in the CK+ database have been provided with a numbered scale from 1 to 5 and hence the target intensity values in the range levels of A – E are accordingly scaled. These AU intensity and expression data are used only for the AU-Emotion Relationship analysis and discriminative AU Set selection, as detailed in Section 3.3.
2. **The Bosphorus 3D Database** includes a rich set of 4652 3D facial scans and corresponding manually labeled facial landmarks collected from 105 subjects (including 60 men and 45 women; 29 of them are professional actors/actresses). Both Action Units (25 out of the 44 defined in FACS) and the six basic emotions are annotated specifically for the purposes of facial expression analysis. The 3D facial scans are acquired by Inspeck Mega Capturor II 3D, with about 0.3mm depth resolution in x , y , and z dimensions and 1600x1200 pixels high color texture resolution (Savran et al., 2008). In this study, excluding occlusion facial scans, a subset of the database containing

clear annotation for both AU intensity and the six basic emotions is considered. The subset includes 729 facial scans covering 56 subjects, and we extract a total of 960 samples for the evaluation of the intensity estimation of the 16 AUs (a scan can contain more than one AUs). These scans contain both frontal and non-frontal head poses with yaw rotations from 0 to 30 degrees and pitch rotations ranging from slight upwards, neutral, to slight downwards.

4.2 Off-line Evaluation

In off-line evaluation, we assess the system's performance by using database sample images with AU intensity and emotion annotations. All the results are obtained using the cross-validation technique. The setting of the off-line evaluation is described in the following.

- For the off-line evaluation, both the training and testing phases were purely based on database images. Therefore, we did not use the Kinect for this evaluation.
- We apply n -fold cross-validation to evaluate the performance of both AU intensity estimation and emotion classification, which embeds training and testing phases of the system together. As detailed in Section 3.2, the cross-validation process uses $n-1$ groups of the data for training and the remaining group for testing. This process is repeated n times. There are overall 729 FACS coded emotional facial images across 56 subjects borrowed from the Bosphorus 3D Database employed for the cross-validation evaluation for both AU intensity estimation and emotion classification. Specifically, we employ 5-fold cross-validation in our work according to the sample size.
- The computational cost of the learning stage in each round of the cross-validation process is approximately 2-5 seconds for AU intensity estimators on average, and 4-6 seconds for emotion classifiers (such as ensemble classifiers) on average. The computational cost of the test stage in each round of cross-validation process is approximately 100-200 milliseconds.

4.2.1 Evaluation on AU Intensity Estimation

As mentioned before, a total of 729 FACS coded emotional facial scans across 56 subjects extracted from the Bosphorus 3D Database (Savran et al., 2008) is used for the evaluation of AU intensity estimation and subsequent emotion classification. The features we used for AU intensity estimation are solely based on the differences of the extracted Euclidean distance features between the neutral and any expressive frames. They are either generated by the manual selection or the mRMR based optimization. For each AU, we have collected around 60 samples, covering both positive cases, i.e. AU presence at any intensity levels (approximately 75%) and negative cases, i.e. AU absence (approximately 25%). A single output value ranging from 0 to 1 is used to represent AU absence through maximum intensity. We apply the 5-fold cross-validation as described above to evaluate the prediction accuracy and generalization capability for each AU. The output AU intensities are subsequently compared against the ground truth to calculate the MSE and CORR for each AU.

In the existing research of AU recognition, the accuracy tends to heavily depend on the training sample size. Typically, most of them required a large number of training images (e.g. thou-

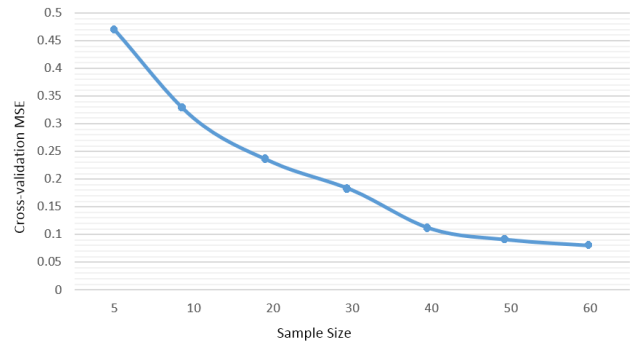


Figure 15. Average cross-validation MSE for AU regression in relation to the data sample size used

sands) with good diversity and coverage to maintain sufficient accuracy and robustness (e.g. Koelstra et al., 2010; Whitehill et al., 2011; Savran et al., 2012). In order to deal with such challenges, we employ the most discriminative motion-based facial features which enable a significant reduction of training data for AU intensity estimation and in the meantime provide an impressive performance. As shown in Figure 15, the average MSE for SVR based AU intensity estimation remains stably below 0.1 once the sample size reaches approximately 50.

Using manually selected features

First, Table 5 shows the results obtained by the feedforward Neural Networks (BPNNs) and Support Vector Regressors (SVRs) for AU intensity estimation using manually selected features. For both BPNNs and SVRs, the lowest MSEs (below 0.05) are observed for AU13 (Cheek Puffer), AU2 (Outer Brow Raiser), AU26 (Jaw Drop), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller) and AU17 (Chin Raiser) followed by AU1 (Inner Brow Raiser), AU15 (Lip Corner Depressor), AU20 (Lip Stretcher), AU18 (Lip Pucker), AU4 (Brow Lowerer), AU23 (Lip Tightner) and AU27 (Mouth Stretch), which also obtain fairly low MSEs below 0.1. These results demonstrate the effectiveness and robustness of the extracted motion-based facial features for AU intensity regression.

In contrast, relative higher MSE (above 0.1) are also ob-

TABLE 5
RESULTS FOR AU INTENSITY ESTIMATION USING MANUALLY SELECTED FEATURES (BPNN= BACKPROPAGATION NEURAL NETWORK, SVR=SUPPORT VECTOR REGRESSION)

AUs	MSE		CORR	
	BPNN	SVR	BPNN	SVR
AU 13	0.011	0.020	0.952	0.957
AU 2	0.013	0.027	0.970	0.978
AU 26	0.025	0.031	0.954	0.976
AU 10	0.036	0.033	0.924	0.939
AU 12	0.042	0.039	0.939	0.930
AU 17	0.043	0.041	0.896	0.923
AU 1	0.047	0.051	0.957	0.960
AU 15	0.056	0.060	0.890	0.892
AU 20	0.058	0.046	0.878	0.913
AU 18	0.064	0.056	0.955	0.947
AU 4	0.066	0.059	0.893	0.824
AU 23	0.092	0.099	0.921	0.925
AU 27	0.097	0.104	0.931	0.969
AU 6	0.119	0.107	0.841	0.859
AU 5	0.134	0.123	0.881	0.895
AU 24	0.149	0.126	0.790	0.863
Overall	0.065	0.063	0.911	0.921

served for the intensity estimation of some AUs, such as AU6 (Cheek Raiser), AU5 (Upper Lid Raiser) and AU24 (Lip Pressor). These results can be explained by the fact that the facial movements of these AUs are very subtle. Especially for AU24, which has the highest MSE and lowest CORR. It could be attributed to the reason that both AU23 and AU24 can cause similar lip boundary changes (e.g. the red parts of lips are narrowed), which may lead to ambiguous annotations even for expert coders. On average, BPNNs and SVRs yield similar performances for AU intensity estimation. However, SVRs are found to perform slightly better than BPNNs for more subtle AUs, in term of both MSE and CORR measurements (e.g. AU5, AU6 and AU24).

Using automatically selected features

Next, we employ the automatically selected features obtained by using the mRMR-based optimization to estimate the intensities of the 16 selected AUs. The results obtained are summarized in Table 6. Empirically, a few informative features with great discrimination power (i.e. 10 to 20 features in general) are sufficient to yield good results. On average, the automatically selected features achieve comparable performance in comparison to the manually selected features for the intensity estimation for many AUs (e.g. AU2, AU13, AU15, AU26, and AU27). For some AUs, such as AU2 and AU13, the automatic features generate even lower MSE values when SVRs are used. However, for some other AUs, such as AU4, AU20 and AU24, the performance drops slightly in comparison to the manual feature selection. Overall, the mRMR-based feature selection yields a very close performance to the manually devised features in terms of both averaged MSE and CORR values. Thus, the AU intensities obtained by SVRs with the corresponding automatically selected features as inputs will be used for subsequent emotion recognition.

Furthermore, since all the results are obtained in the form of continuous AU intensity levels, they reflect more physical truth of facial expressions in comparison to other applications that only performed presence or absence binary-classifications (e.g. Tsalakanidou & Malassiotis, 2010; Li et al., 2013). Such AU intensity measurements may also indicate effective physical

cues to contribute to the sequent emotion classification.

4.2.2 Evaluation on Facial Emotion Recognition

The 729 facial scans used for AU intensity estimation above are then applied for the evaluation of the facial emotion recognition. As mentioned before, the intensities of the 16 diagnostic AUs generated by SVRs with mRMR based feature selection are subsequently used as inputs to the six ensemble classifiers for expression recognition. Six single SVM classifiers are also used to perform facial expression recognition for the comparison with the ensemble classifiers. We also apply a 5-fold cross-validation to measure the accuracy performance of each emotion recognition classifier. We measure the performance of the proposed emotion recognition approaches in term of the accuracy confusion matrix and F1-measure. A confusion matrix is a $n \times n$ matrix, where the row labels are ground-truth emotion annotations and the column labels are the classification results. The diagonal entries indicate the correct classifications, while the off-diagonal entries correspond to misclassifications. The F1-measure is a harmonic mean of precision and recall rate, which is considered to be a more comprehensive metric.

Table 7 shows the recognition accuracy confusion matrices for the six basic emotions obtained by SVMs and the proposed ensemble classifiers. By using SVMs for emotion classification, we achieve an overall recognition accuracy rate of 90.5% (shown in Table 7 (a)), while by using ensemble models, we obtain a higher overall accuracy of 92.2% (see Table 7 (b)). More specifically, for either approach, the best performances are achieved for the recognition of ‘happy’ and ‘surprised’ facial expressions, with recognition accuracies beyond 95%. For ‘anger’ and ‘fear’, slightly lower recognition accuracies are observed for both approaches with the ensembles (92.8% for ‘anger’ and 92.1% for ‘fear’) outperforming the SVM classifiers (91.3% for ‘anger’ and 91.1% for ‘fear’). For ‘disgust’, a lower recognition accuracy of 85.6% is observed when using the SVMs, and 88.6% when using the ensembles. A possible explanation is that those emotions with comparatively lower recognition accuracies often entangled with more complicated and subtle facial changes than the ones with higher recognition accuracies, and thus more challenging to recognize. The lowest recognition rates are observed for ‘sadness’ (82.7% by SVM and 86.6% by the ensemble classifier). This could be due to the fact that in some facial scans, subjects inaccurately express ‘sadness’ using the combination of AU20 (Lip Stretcher) and AU15 (Lip Corner Depressor), rather than solely using AU15 as indicated by FACS (Ekman et al., 2002). But AU20 is also served as a key physical cue for ‘fear’, which may lead to misclassification of ‘sadness’ as ‘fear’.

We subsequently compare our work with other state-of-the-art developments such as Salahshoor & Faez (2012) and Ujir (2013) in Table 8. These related applications are chosen because of their focus on a similar research challenge of 3D facial emotion recognition and the employment of the same Bosphorus 3D database and similar evaluation strategies. Salahshoor & Faez (2012) proposed a novel dynamic mask to automatically segment the regions of face which were less sensitive to expressions and applied a modified nearest neighbor classifier for the recognition of the six basic emotions. Moreover, Ujir (2013) decomposed a face into six distinct regions and extracted their 3D facial surface normals instead of raw 3D points as the fea-

TABLE 6

RESULTS FOR AU INTENSITY ESTIMATION USING AUTOMATICALLY SELECTED FEATURES (BPNN= BACKPROPAGATION NEURAL NETWORK, SVR=SUPPORT VECTOR REGRESSION)

AUs	MSE		CORR	
	BPNN	SVR	BPNN	SVR
AU 2	0.013	0.017	0.937	0.953
AU 13	0.021	0.014	0.919	0.975
AU 26	0.032	0.031	0.923	0.975
AU 10	0.039	0.041	0.885	0.938
AU 12	0.059	0.053	0.895	0.926
AU 17	0.057	0.059	0.873	0.900
AU 1	0.066	0.060	0.906	0.936
AU 15	0.066	0.062	0.874	0.891
AU 20	0.059	0.064	0.875	0.912
AU 18	0.077	0.069	0.911	0.936
AU 4	0.080	0.078	0.897	0.805
AU 23	0.095	0.094	0.893	0.905
AU 27	0.102	0.097	0.886	0.963
AU 6	0.120	0.117	0.822	0.838
AU 5	0.136	0.133	0.831	0.878
AU 24	0.152	0.142	0.787	0.857
Overall	0.073	0.071	0.882	0.912

TABLE 7
CONFUSION MATRICES OF FACIAL EMOTION RECOGNITION ACCURACIES

	<i>Anger</i>	<i>Disgust</i>	<i>Fear</i>	<i>Happy</i>	<i>Sadness</i>	<i>Surprise</i>
<i>a. Recognition accuracy (average 90.5%) using SVM classifiers</i>						
<i>Anger</i>	91.3	6.7	0	0	13.6	0
<i>Disgust</i>	11.1	85.6	4.3	0	0	0
<i>Fear</i>	0	0	91.1	0	0	11.8
<i>Happy</i>	0	0	0	95.6	0	8.8
<i>Sadness</i>	5.8	3.1	9.8	0	82.7	0
<i>Surprise</i>	0	0	0	7.9	0	96.5
<i>b. Recognition accuracy (average 92.2%) using the proposed ensemble classifiers</i>						
<i>Anger</i>	92.8	3.3	0	0	9.3	0
<i>Disgust</i>	9.8	88.6	2.3	0	0	0
<i>Fear</i>	0	0	92.1	0	0	9.9
<i>Happy</i>	0	0	0	96.1	0	8.9
<i>Sadness</i>	4.7	0	7.3	0	86.6	0
<i>Surprise</i>	0	0	0	7.3	0	96.7

ture vectors. Then Support Vector Machines were employed to recognize facial expressions for the six regions independently. A weighted voting scheme was also applied to make the final classification. The comparison in Table 8 indicates that our proposed facial emotion recognition system outperforms both of the above related developments. Specifically, the ‘surprised’ facial expression has been well recognized by all the three systems (accuracies > 90%). However, the two related systems also respectively show some limitations for the recognition of some of the other emotion categories. For example, the system of Salahshoor & Faez (2012) performed poorly for the recognition of ‘happy’ and ‘disgust’ (accuracies < 80%) emotions, whereas the work of Ujir (2013) also indicated very unstable classification performance for ‘fear’ (only 21.5%) and ‘disgust’ (43.1%)

TABLE 8
COMPARISON OF RECOGNITION ACCURACIES FOR THE SIX BASIC EMOTIONS

	<i>Accuracy _SVM</i>	<i>Accuracy _Ensemble</i>	Salahshoor & Faez (2012)	Ujir (2013)
<i>Surprise</i>	96.5	96.7	91.4	90.8
<i>Happy</i>	95.6	96.1	74.3	100.0
<i>Fear</i>	91.1	92.1	92.9	21.5
<i>Anger</i>	91.3	92.8	87.3	75.4
<i>Disgust</i>	85.6	88.6	78.3	43.1
<i>Sadness</i>	82.7	86.6	95.5	67.7
<i>Overall</i>	90.5%	92.2%	86%	66.4%

TABLE 9
F1-MEASURES FOR THE SIX BASIC EMOTIONS

	<i>F1_SVM</i>	<i>F1_Ensemble</i>	<i>F1_Sandbach et al. (2012)</i>
<i>Surprise</i>	0.889	0.897	0.826
<i>Happy</i>	0.94	0.945	0.812
<i>Fear</i>	0.888	0.913	0.462
<i>Anger</i>	0.877	0.895	0.500
<i>Disgust</i>	0.876	0.923	0.644
<i>Sadness</i>	0.843	0.884	0.625
<i>Overall</i>	0.89	0.91	0.65

expressions. In comparison to these state-of-the-art applications, our system is proved to be more stable for the recognition of all of the six emotion categories and achieves the highest overall recognition accuracy among the related applications.

Since the classification accuracy rate could be less informative sometimes, especially when the data is unbalanced, the F1-measure for each emotion category is also presented in Table 9. We also compare our system with the work by Sandbach et al. (2012) because of their state-of-the-art performance and the employment of the same performance metric (i.e. the F1-measure). In their work, hidden Markov models (HMMs) were used to recognize the six basic emotions from facial expressions based on 3D modality. F1-measure was also produced for each emotion category. Based on the comparison of the F1-measure results, it is noticed that the performance of our system significantly outperforms those of the work by Sandbach et al. (2012). Although their HMM based approach also generated good results for the recognition of ‘happy’ and ‘surprised’ facial expressions, our system performs more stably for the detection of each emotion category. Overall, the above results demonstrate that the proposed system is consistently an efficient and robust solution for AU intensity estimation and emotion recognition.

Furthermore, facial expressions sometimes may contain a mixture of emotions, thus it is possible that two (or more) emotional states occur simultaneously in one emotional facial scan. This research also shows great potential to detect such combination of emotions (e.g. happy + surprise) by deriving recognition results for each emotion category separately.

4.3 On-line Evaluation

The facial emotion recognition system has also been applied to real-time emotion detection tasks contributed by test human subjects. The facial feature point localization of our system is able to integrate both color and 3D depth image data so that it provides great robustness against illumination changes and pose variations. It thus lays solid foundations for subsequent AU intensity measurement and emotion recognition. Moreover, the computational complexity of the face tracking and landmark localization requires 20-30 milliseconds under normal lab lighting conditions. The mRMR-based feature selection, AU regression, and emotion classification take an averaged run time of 3-5 milliseconds (which may change slightly depending on different types of regressors and classifiers used). Overall, the system

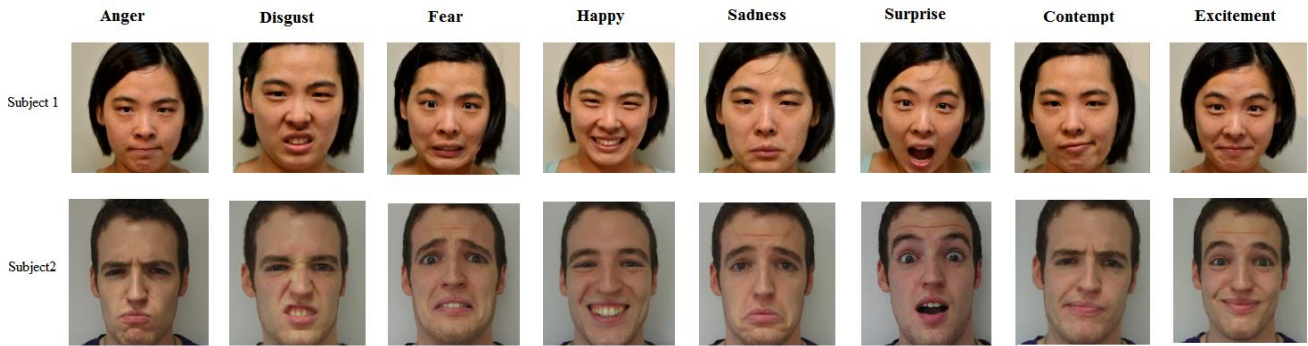


Figure 16. Snapshots of the six basic emotions plus ‘contempt’ and ‘excitement’ posed by two test subjects in the on-line evaluation

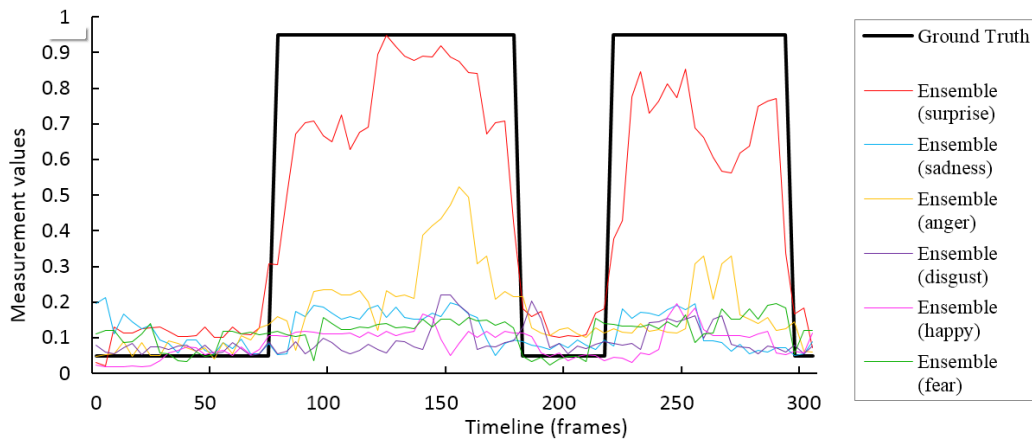


Figure 17. Examples of real-time detection of ‘surprise’. The bold black line indicates the ground-truth (presence/absence), and the six color lines respectively indicate the real-time outputs of the six ensemble classifiers

is able to perform efficiently for facial emotion recognition at a frame rate of 25~30 fps on i7 quad-core CPUs with 8GB RAM.

For the on-line evaluation, our system has been trained with database images first and then is used to recognize human subjects’ facial expressions in real time. The setting of the online testing is provided in the following.

- For the online evaluation, our system has been trained with database images first. Then the Kinect is used in the testing phase to track human subjects’ facial landmarks. Based on the tracked facial landmarks, the system subsequently performs mRMR-based feature selection, AU intensity estimation and emotion recognition.
- For the on-line evaluation, the above 729 FACS coded database images from 56 subjects employed for the off-line evaluation are entirely used for training of both the AU intensity estimators and emotion classifiers. For testing, we recruit eleven real human subjects to further evaluate the system’s efficiency.
- In this real-time evaluation, the training computational cost of the system is approximately 4-5 seconds for AU intensity estimators while 5-7 seconds for emotion classifiers. For on-line testing, we recruit eleven participants with five females and six males aging from 25 to 40 years old. Majority of them are postgraduate students and all the test subjects are non-experts in the field. The computational cost of the system in the real-time testing is about 3-5 milliseconds.

As mentioned above, we recruit eleven participants for real-time system testing. In order to ensure effective tracking of

facial geometric features, the distance between the participants and the Kinect was controlled within the range of 2 (± 0.5) meters. The participants were required to display a series of emotional clips. Each clip lasts approximately 10–15 seconds (i.e. 300–450 frames). It starts from a short neutral state period (4–5 seconds) and followed by a posed facial expression period. Both the neutral state and expression periods were manually labeled in each clip by an expert annotator. In addition to the six basic emotions (happiness, sadness, disgust, surprise, fear and anger) that are collected from the test subjects and used to test the system, we also evaluate the system with some novel emotional expressions (e.g. contempt and excitement) contributed by the test subjects.

In our experiment, the expressions of ‘contempt’ emotion require a subject to show the facial behavior of dimpler (AU14) while the expressions of ‘excitement’ emotion require the combination of ‘surprise’ and ‘happy’ expressions with the upper face showing inner and outer brow raiser and upper lid raiser and the lower face indicating cheek raiser and lip corner puller. We use the above guidance for the posing and collection of these two novel emotion classes for testing. Figure 16 shows examples of the six basic emotions plus ‘contempt’ and ‘excitement’ expressions posed by two test subjects during testing. Eventually, the system was evaluated with a total of 136 emotional clips. The detailed results and discussions are presented as follows.

Figure 17 shows an example of real-time detection of a ‘surprise’ emotional clip using the six ensemble classifiers. The

TABLE 10
REAL-TIME RECOGNITION ACCURACIES FOR THE SIX BASIC
EMOTIONS AND NOVEL EMOTION CLASSES

	<i>Recognition Accuracy (average 84%)</i>
<i>Surprise</i>	93.2
<i>Happy</i>	88.1
<i>Fear</i>	81.6
<i>Anger</i>	79.4
<i>Disgust</i>	83.7
<i>Sadness</i>	77.9
	<i>Classified as a novel emotion (average 72.2%)</i>
<i>Contempt</i>	77.2
<i>Excitement</i>	67.1

vertical axis indicates the emotion detection results from absence (0) to maximum presence (1) of the ‘surprise’ expression, and the horizontal axis marks the timeline (in frames). As illustrated in Figure 17, for the recognition of ‘surprise’, ideally, only the corresponding ensemble classifier for ‘surprise’ generates an output curve consistent with the ground truth. The outputs of the other five ensemble classifiers consistently remain in a much lower level. Overall, the average classification accuracy rate for this emotion clip is 93.2%.

Table 10 summarizes the real-time recognition accuracy rates for the six basic emotions and novel emotion detection rates for ‘contempt’ and ‘excitement’. Generally, the on-line system yields comparable results to that were obtained in off-line evaluation. Except for ‘anger’ and ‘sadness’, the recognition accuracy rates for the other four basic emotions are consistently beyond 80%. Moreover, 77.2% of ‘contempt’ and 67.1% of ‘excitement’ expressions are successfully identified as novel emotion classes, which demonstrate that the proposed ensemble classifiers are well capable of detecting newly arrived novel emotion categories.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented a fully automatic system for real-time 3D AU intensity estimation and emotion recognition. We first realized real-time 3D face tracking and facial landmark extraction based on the Kinect platform. Then 16 sets of motion-based facial features containing rich person-independent emotional information were extracted and selected by using both manual and mRMR-based automatic feature selection methods. These feature sets were subsequently employed as inputs to an array of Neural Networks and Support Vector Regressors respectively to estimate the intensities of the 16 diagnostic AUs. Experimental results indicated that the mRMR based optimized feature selection yields comparable results in comparison to the manually selected features when using either Neural Networks or SVRs for AU intensity measurement. Moreover, the SVR-based AU intensity estimation slightly outperformed the Neural Network based method. This is probably caused by the fact that the grid search with cross validation has been conducted for optimal parameter selection for the SVR models. By using the automatically selected features and SVRs, we have achieved an averaged MSE of 0.071 and an averaged

CORR of 0.912 for the intensity estimation of the 16 AUs. The intensities of AU2 (Outer Brow Raiser), AU10 (Upper Lip Raiser), AU13 (Cheek Puffer) and AU26 (Jaw Drop) were well estimated with lowest errors ($MSE < 0.05$), whereas more subtle AUs, such as AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), and AU24 (Lip Pressor) were estimated with relatively higher estimation errors ($MSE > 0.1$). The above results also demonstrated the extracted motion-based facial features are very efficient and robust for AU intensity estimation.

We subsequently used the derived AU intensities to recognize the six basic emotions using the identified discriminative AU combinations and dedicated ensemble classifiers for each emotion category. The proposed novel adaptive ensemble classifiers show great robustness and flexibility for not only the recognition of six basic emotions but also the detection of newly arrived unseen novel emotion categories. The off-line evaluation results using the Bosphorus database indicated that the proposed ensemble models consistently outperform the SVM-based classification, and have achieved an averaged recognition accuracy of 92.2% and an averaged F1-measure of 91% for the recognition of the six basic emotions. The best recognition accuracies were obtained for ‘happy’ and ‘surprise’ facial expressions ($> 96\%$) with ‘fear’, ‘anger’ and ‘disgust’ reasonably recognized ($> 88\%$). The lowest recognition accuracy rate was observed for ‘sadness’ (86.6%). The system also outperforms other state-of-the-art research on 3D facial emotion recognition tasks based on the comparison of both the recognition accuracy and F1-measure results.

We also conducted an on-line evaluation with real human subjects to assess the system’s real-time performance and the efficiency for novel emotion class detection. Overall, the proposed system is able to perform facial emotion recognition efficiently with a frame rate of 25~30 fps on i7 quad-core CPUs with 8GB RAM. We obtained an impressive average recognition accuracy rate of 84% for the detection of the six expressions when tested with real human subjects (only slightly lower than those achieved in off-line evaluation). Moreover, the proposed ensemble classifiers also show superior ability to detect the arrival of novel emotion classes with 72.2% detection rate on average.

In future work, the facial anatomy and FACS domain knowledge that closely related to facial muscle movements and subtle facial expressions will be further studied so that we can identify more effective dynamic facial features to recognize a wider variety of emotions, especially compound emotions (e.g. happy surprise and angry surprise). We will also further validate the system’s performance using more challenging spontaneous facial expressions in real-life interactions, since in such spontaneous expressions, AUs usually occur with relatively lower intensities in more subtle combinations comparing to the posed ones. Furthermore, other state-of-the-art 3D facial image databases and layered cascade optimization techniques for feature dimensionality will also be employed to further improve the robustness and efficiency of the proposed system. Finally, we also aim to incorporate each weak affect indicator embedded in body language (e.g. gestures) with emotional facial expression recognition to draw more reliable affect interpretation. We believe these are crucial aspects for the development of personalized effective human-like agent-based interfaces.

ACKNOWLEDGMENT

The authors appreciate suggestions and comments from the reviewers and the time the reviewers spent for the review of our paper.

REFERENCES

- Ahlberg, J. (2001). CANDIDE-3—an updated parameterized face. Report No. LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden.
- Antonini, G., Sorci, M., Bierlaire, M., & Thiran, J. (2006). Discrete choice models for static facial expression recognition. *8th International Conference on Advanced Concepts for Intelligent Vision Systems*, Lecture Notes in Computer Science, vol. 41, Springer, Berlin, pp.710-721.
- Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., & Movellan, J.R. (2006). Automatic recognition of facial actions in spontaneous expressions. *J. Multimed*, 1(6), 22-35.
- Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing - Letters and Review*, 11(10).
- Besinger, A., Sztynka, T., Lal, S., Duthoit, C., Agbinya, J., Jap, B., Eager, D., & Dissanayake, G. (2010). Optical flow based analyses to detect emotion from human facial image data. *Expert Systems with Applications*, 37, 8897–8902.
- Bouquet, Y. L. (1999). Pyramidal implementation of the Lucas–Kanade feature tracker. *Technical Report*, Intel Corporation, Microprocessor Research Labs.
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27: 1-27:27, 2011.
- Chang, K.Y., Liu, T.L., & Lai, S.H. (2009). Learning partially-observed hidden conditional random fields for facial expression recognition. *Computer Vision and Pattern Recognition, CVPR 2009. IEEE Conference on*, 533-540.
- Chang, Y., Hu, C., & Turk, M. (2004). Probabilistic Expression Analysis on Manifolds. In *Proc. IEEE International 'I Conf. Computing. Vis. Pattern Recognition*.
- Chavan U. B., & Kulkarni D. B. (2013). Facial Expression Recognition-Review. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, Vol. 3, No. 1, pp. 237-243.
- Cheng, S. C., Chen, M. Y., Chang, H. Y., & Chou, T. C. (2007). Semantic-based facial expression recognition using analytical hierarchy process. *Expert Systems with Applications*, 33(1), 86–95.
- Cohn, J., Kruez, T.S., Matthew, I., Yang, Y., Nguyen, M., Padilla, M.T., Zhou, F., & Torre, F.D. (2009). Detecting depression from facial actions and vocal prosody. In *International Conference on Affective Computing and Intelligent Interaction (ACII2009)*.
- D'Mello, S., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language and facial features. *User Model User-Adapt Interact*, 20(2), 147-187.
- Davis, R.A., Charlton, A., Oehlschlager, S., & Wilson, J. (2006). Novel feature selection method for genetic programming using metabolomic ¹H NMR data. *Chemometrics and Intelligent Laboratory Systems*, 81(1), 50-59.
- DeGroot, M. H., & Schervish, M. J. (2011). *Probability and Statistics* (4th edition). Published by Pearson.
- Ekman, P., & Friesen, W.V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124-129.
- Ekman, P., & Friesen, W.V. (1983). *Emfacs-7: Emotional Facial Action Coding System*. University of California at San Francisco.
- Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial Action Coding System, the Manual*. Published by Research Nexus division of Network Information Research Corporation, USA.
- Ekman, P., Friesen, W.V., & Hager, J.C. (2002). Facial Action Coding System Investigator's Guide. *Consulting Psychologist Press*, Palo Alto, CA.
- Farid, D., Zhang, L., Hossain, A.M., Rahman, C.M., Strachan, R., Sexton, G., & Dahal, K. (2013). An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams. *Expert Systems with Applications*, Vol 40, Issue 15. 5895-5906.
- G'Mussel, A.S., & Hewig, J. (2013). The value of a smile: Facial expression affects ultimatum-game responses. *Judgment and Decision Making*, 8 (3), 381-385.
- García-Pedrajas, N., Hervás-Martínez, C., & Ortiz-Boyer, D. (2005). Cooperative Coevolution of Artificial Neural Network Ensembles for Pattern Classification. *IEEE Transactions on Evolutionary Computation*, 9(3), 271–302.
- Gu, W., Xiang, C., Venkatesh, Y.V., Huang, D., & Lin, H. (2012). Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, 45, pp. 80-91.
- Hecht-Nielsen, R. (1989). Theory of the backpropagation neural network. *Neural Networks, 1989. IJCNN, International Joint Conference on*, San Diego, CA, USA.
- Hsu, C., Chang, C., & Lin, C. (2010). *A practical guide to support vector classification*. Department of Computer Science National, Taiwan University.
- Jeatrakul, P. & Wong, K.W. (2009). Comparing the performance of different neural networks for binary classification problems. *Natural Language Processing, SNLP '09. Eighth International Symposium on*, 111-115.
- Jong, D.H., Ziemkiewicz, C., Ribarsky, W., and Chang, R. (2009). Understanding Principal Component Analysis Using a Visual Analytics Tool. *Charlotte Visualization Center, UNC Charlotte*.
- Kaltwang, S., Rudovic, O., & Pantic, M. (2012). Continuous pain intensity estimation from facial expressions. in *Advances in Visual Computing. Lecture Notes in Computer Science*, vol. 7432. Heidelberg: Springer, 368–377.
- Kappas, A. (2010). Smile When You Read This, whether you like it or not: Conceptual Challenges to Affect Detection. *Affective Computing, IEEE Transactions on*, 1(1), 38-42.
- Koelstra, S., Pantic, M., & Patras, I. (2010). A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Transactions on, Pattern Analysis and Machine Intelligence*, 32 (11), 1940-1954.
- Kotsia, I., Zafeiriou, S., & Pitas, I. (2008). Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition*, 41, 822-851.
- Kraipeerapun, P. (2008). Neural network classification based on quantification of uncertainty, Murdoch University.
- Li, Y., Chen, J., Zhao, Y., & Ji, Q. (2013). Data-free Prior Model for Facial Action Unit Recognition. *Affective Computing, IEEE Transactions on*, early accepted.
- Lucey, P., Cohn, J., Lucey, S., Matthews, I., Sridharan, S., & Prkachin, K. (2009). Automatically Detecting Pain Using Facial Actions. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 1-8.
- Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, In *Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis*, San Francisco, USA.
- Lucey, S., Matthews, I., Hu, C., Cohn, J., & Ambadar, Z. (2006). AAM derived face representations for robust facial action recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*.
- Mahoor, M. H., Zhou, M., Veon, S., & Cohn, J. (2011). Facial action unit recognition with sparse representation. In *Proc. IEEE Automatic Face & Gesture Recognition and Workshops (FG 2011)*.
- Masud, M.M., Gao, J., Khan, L., Han, J., & Thuraisingham, B. (2011). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23, 859–874.
- Metallinou, A., Katsamanis, A., & Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2), 137-152.
- Mitchell, T., & Hill, M. (1997). *Machine Learning*. Publisher McGraw-Hill, Inc., New York, USA.
- Mpiperis, L. (2008). 3D facial expression recognition using swarm intelligence. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Las Vegas, Nevada, USA.

- Owusu, E., Zhan, Y., & Mao, Q.R. (2014). A neural-AdaBoost based facial expression recognition system. *Expert Systems with Applications*, 41(7), 3383-3390.
- Pandzic, I., & Forchheimer, R. (2002). *MPEG-4 Facial Animation: the Standard, Implementation and Applications*. Wiley.
- Pantic, M., & Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man and Cybernetics*, Part B: Cybernetics, 36 (2), 433-449.
- Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 27, 1226-1238.
- Rao, K.S., Saroj, V.K., Maity, S., & Koolagudi, S.G. (2011). Recognition of emotions from video using neural network models. *Expert Systems with Applications*, 38(10), 13181-13185.
- Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, Volume 33, Issue 1-2, 1-39.
- Ryan, A., Cohn, J., Lucey, S., Saragih, J., Lucey, P., & Rossi, A. (2009). Automated Facial Expression Recognition System. In *Proceedings of the International Carnahan Conference on Security Technology*, 172-177.
- Salahshoor, S. & Faez, K. (2012). 3D Face Recognition Using an Expression Insensitive Dynamic Mask. *Image and Signal Processing, lecture Notes in Computer Science*, Volume 7340, 253-260.
- Sandbach, G., Zafeiriou, S., Pantic, M., & Rueckert, D. (2012). Recognition of 3D facial expression dynamics. *Image Vision Computing*, issue 30, 762-773.
- Savran, A., Alyuz, N., Dibeklioglu, H., Celiktutan, O., Gokberk, B., Sankur, B., & Akarun, L. (2008). Bosphorus database for 3D face analysis. In: *Proc. First COST 2101 Workshop on Biometrics and Identity Management*, Denmark, 47-56.
- Savran, A., Sankur, B., & Bilge, M.T. (2012). Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units. *Pattern Recognition*, volume 45, pp.767-782.
- Savran, A., Sankur, B., & Bilge, M.T. (2012). Regression-based intensity estimation of facial action units. *Pattern Recognition*, volume 30, 774-784.
- Shan, C., Gong, S., & McOwan, P.W. (2009). Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study. *Image and Vision Computing*, vol. 27, 803-816.
- Sikora, R., & Piramuthu, S. (2007) Framework for efficient feature selection in genetic algorithm based data mining. *European Journal of Operational Research*, 180(2), 723-737.
- Sorci, M., & Thiran, J.Ph. (2010). Modeling human perception of static facial expressions. *Image and Vision Computing*, 28 (5). 790-806.
- Soyel, H., & Demirel, H. (2007). Facial Expression Recognition Using 3D Facial Feature Distances. *ICIA07*, 831-838.
- Srivastava, R., & Roy, S. (2009). 3D facial expression recognition using residues. In *TENCON, 2009-2009 IEEE Region 10 Conference*, 1-5.
- Swets D.L., & Weng, J. (1996). Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, pp. 831-836.
- Tang, H., & Huang, T. S. (2008). 3D facial expression recognition based on properties of line segments connecting facial feature points. In *Proc. 8th IEEE International Conf. Automatic Face Gesture Recognition*, 1-6.
- Tang, H., & Huang, T. S. (2008). 3D facial expression recognition based on automatically selected features. *Computer Vision and Pattern Recognition Workshops, IEEE Computer Society Conference on*, 1-8.
- Tian, Y. L. (2002). Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FG'02)*. Washington, DC.
- Tong, Y., Chen, J., & Ji, Q. (2010). A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 32 (2), 258-273.
- Tong, Y., Liao, E., & Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10), 1683-1699.
- Torralba, A., & Efros, A. (2011). Unbiased Look at Dataset Bias. *IEEE Conference on Computer Vision and Pattern Recognition*.
- Tsalakanidou, F., & Malassiotis, S. (2010). Real-time 2D+3d Facial Action and Expression Recognition. *Pattern Recognition*, 43 (5), 1763-1775.
- Ujir, H. (2013). 3D facial expression classification using a statistical model of surface normals and a modular approach. *Ph.D. thesis, University of Birmingham*.
- Valstar, M., & Pantic, M. (2006). Biologically vs. logic inspired encoding of facial actions and emotions. In *Proc. of IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 325-328.
- Valstar, M., & Pantic, M. (2007). Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics. *Lecture Notes on Computer Science*, vol. 4796, 118-127.
- Valstar, M., & Pantic, M. (2012). Fully Automatic Recognition of the Temporal Phases of Facial Actions. *IEEE Transactions on Systems, Man, and Cybernetics*, Part B: Cybernetics, 42 (1), 28-43.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, USA.
- Vapnik, V. N. (2001). *The nature of Statistical learning theory*, 2nd edition. Springer, New York, USA.
- Vukadinovic, D., & Pantic, M. (2005). Fully automatic facial feature point detection using Gabor feature based boosted features. In *Proc. IEEE Int. Conf. Syst., Man, and Cybern*, 1692-1698.
- Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., & Movellan, J. (2008). Automated Drowsiness Detection for Improved Driver Safety. In *Proceedings of the International Conference on Automotive Technologies*, 2008.
- Wang, J., Yin, L., Wei, X., & Sun, Y. (2006). 3D facial expression recognition based on primitive surface feature distribution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA.
- Wang, T. H., & Lien, J. (2009). Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation. *Pattern Recognition*, vol. 42, 962-977.
- Wang, X., Yang, J., Teng, X., Xia, W., & Jensen, R. (2007). Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters*, 28(4), 459-471.
- Webb, J., & Ashley, J. (2012). *Beginning Kinect programming with the Microsoft Kinect SDK*. USA, Published by Apress.
- Wen, Z., & Huang, T. (2003). Capturing subtle facial motions in 3d face tracking. In *Proc. of Int. Conf on Computer Vision*.
- Whitehill, J., & Omlin, C.W. (2006). Haar Features for FACS AU Recognition. *Automatic Face and Gesture Recognition, FGR 2006. 7th International Conference on*, 217-222.
- Whitehill, J., Tingfan, W., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). *IEEE Conference on, Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 298-305.
- Zavaschi, T., Oliveira, L.S., Souza Jr, A.B., and Koerich, A. (2013). Fusion of feature sets and classifiers for facial expression recognition. *Expert System with Applications*, 40(2), 646-655.
- Zhang, L. (2011). Facial Expression Recognition Using Facial Movement Features. *Affective Computing, IEEE Transactions on*, 2 (4), 219-230.
- Zhang, L., Jiang, M., Farid, D., & Hossain, M.A. (2013). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications*, 40(13), 5160-5168.